ACL 2025 VIENNA

JULY 27 - AUGUST 1

The 63rd Annual Meeting
of the Association for
Computational Linguistics



Contents

Table of Contents **i**

1 Conference Information **1**
General Chair Message . . . 1
Program Chair Message . . . 4
ACL Business and Event Manager Message . . . 8
Organizing Committee . . . 9
Program Committee . . . 12

2 Anti-harassment Policy **17**
Ethics Policy . . . 17

3 Meal Info **19**

4 Social Events **23**
ACL 2025 Welcome Reception . . . 23
ACL 2025 Findings Reception . . . 24
ACL 2025 Social Event . . . 25

5 Keynotes **27**
Monday, July 28 . . . 27
Tuesday, July 29 . . . 29

6 Panel Discussion **31**
Monday, July 28 . . . 31

7 Tutorials **35**
Overview . . . 35
Turorial 1 . . . 37
Turorial 2 . . . 38

Turorial 3 . . . 39
Turorial 4 . . . 40
Turorial 5 . . . 41
Turorial 6 . . . 42
Turorial 7 . . . 43
Turorial 8 . . . 44

8 Main Conference **45**
Main Conference: Monday, July 28, 2025 . . . 45
Main Conference: Tuesday, July 29, 2025 . . . 47
Main Conference: Wednesday, July 30, 2025 . . . 49

9 Day 1: Monday 28 July **51**
Oral Presentation . . . 51
Poster Presentation . . . 74
Virtual Presentation . . . 218

10 Day 2: Tuesday 29 July **283**
Oral Presentation . . . 283
Poster Presentation . . . 303
Virtual Presentation . . . 583

11 Day 3: Wednesday 30 July **647**
Oral Presentation . . . 647
Poster Presentation . . . 660
Virtual Presentation . . . 799

12 Workshops **815**
Overview . . . 815
W01: 6th AfricaNLP Workshop: Multilingual and Multicultural-aware LLMs . . . 817
W02: The 12th Workshop on Argument Mining . . . 818
W03: 2nd Workshop on Natural Language Processing meets Climate Change . . . 819
W04: Eighth Workshop on Fact Extraction and VERification (FEVER) . . . 820
W05: Workshop on Meaningful, Efficient, and Robust Evaluation of LLMs and GEM: Natural Language Generation, Evaluation, and Metrics . . . 821
W06: The 19th Linguistic Annotation Workshop (LAW XIX) . . . 822
W07: The 4th Workshop on NLP for Positive Impact . . . 823
W08: REALM: First Workshop for Research on Agent Language Models . . . 824
W09: SDP 2025: The 5th Workshop on Scholarly Document Processing . . . 825
W10: The Third Workshop on Social Influence in Conversations (SICon 2025) . . . 826
W11: Slavic NLP-2025: The 10th Biennial Workshop on Slavic NLP . . . 827
W12: The 4th Table Representation Learning Workshop at ACL 2025 . . . 828
W13: The Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025) . . . 829
W14: The 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025) . . . 830
W15: 29th Conference on Computational Natural Language Learning (CoNLL) . . . 831
W16: The 22nd International Conference on Spoken Language Translation (IWSLT 2025) . . . 832
W17: The 19th Workshop on Semantic Evaluation (Semeval-2025) . . . 833
W18: The Second Workshop on Analogical Abstraction in Cognition, Perception, and Language (Analogy-ANGLE II) . . . 834
W19: BioNLP 2025 and Shared Tasks (BioNLP-ST 2025) . . . 835
W20: The 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP 2025) . . . 836
W21: The First Workshop on Large Language Model Memorization (L2M2) . . . 837

W22: First Workshop on LLM Security (LLMSEC) . . . 838
W23: MAGMaR: Multimodal Augmented Generation via MultimodAl Retrieval . . . 839
W24: FieldMatters–SIGTYP: The 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP) and the Fourth Workshop on NLP Applications to Field Linguistics (FieldMatters) . . . 840
W25: The 3rd workshop on Towards Knowledgeable Foundation Models . . . 841
W26: WikiNLP: Advancing Natural Language Processing for Wikipedia . . . 842
W27: The 9th Workshop on Online Abuse and Harms (WOAH 2025) . . . 843
W28: The First Joint Workshop on Large Language Models and Structure Modeling . . . 844

13 Poster Printing **845**

14 Local Guide **847**

15 Venue Map **851**

Sponsorship **855**

1 Conference Information

General Chair Message

It is my great pleasure and honor to welcome you to the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025), held in beautiful Vienna, Austria, from July 27 to August 1, 2025. ACL 2025 continues our field's tradition of excellence in scholarship, innovation, and inclusivity, and I am deeply grateful to the many volunteers who have worked tirelessly to bring this event to life.

I want to express my deepest thanks to our Program Chairs — **Wanxiang Che, Joyce Nakatumba-Nabende, Mohammad Taher Pilehvar** and **Ekaterina Shutova** — who have overseen the reviewing and selection process, and shaped a compelling and diverse scientific program. Overall, we received more than 8300 submissions and accepted 1700 at the main conference and 1392 as findings. This was made possible through close coordination with the ACL Rolling Review (ARR) team, led by the ARR Editors-in-Chief — **Jun Suzuki, Jing Jiang**, and **Xiaodan Zhu** — and I sincerely thank them. Although some improvements are still possible, the ARR workflow was already well integrated with the conference review and decision-making process. We are also deeply grateful to the (many!) Senior Area Chairs, Area Chairs, reviewers, and the Best Paper Committee (led by **Rada Mihalcea** and **Roi Reichart**), whose dedication ensured the high quality of our program. I am also grateful to our Technical Open Review Chairs, **Niket Tandon** and **Lizhen Qu**, for their behind-the-scenes work. For those joining virtually, our Virtual Infrastructure Chairs — **Manling Li, Yang Liu**, and **Avi Sil** — have worked to make the hybrid conference experience inclusive and engaging. I would also like to recognize the efforts of our Ethics Chairs — **Karën Fort** and **Bjorn Ross** — who have led the work aimed at ensuring that submissions meet the ethical standards of our field.

ACL 2025 hosts a rich set of workshops and tutorials, thanks to the dedicated efforts of our Workshop Chairs — **Terra Blevins** and **Christophe Gravier** — and our Tutorial Chairs — **Yuki Arase, David Jurgens**, and **Fei Xia**. Our Demonstration Track Chairs — **Pushkar Mishra, Smaranda Muresan**, and **Tao Yu** — have put together an impressive set of system demonstrations of cutting-edge innovations in NLP. I am proud to say that we also have a rich Industry Track, thanks to **Yunyao Li** and **Georg Rehm** who created a "conference in the conference". Given the growing interaction between academia and industry, I hope this tradition, which stems from NAACL and which I reinstated at ACL, will be continued in the following editions.

As usual, TACL and CL papers are presented at the conference. Thanks go to the TACL Editors-in-Chief

(**Asli Celikyilmaz, Roi Reichart**, and **Dilek Hakkani-Tur**) and the CL Editor-in-Chief, **Wei Lu**, for their coordination efforts. I am also proud to announce that this year we are introducing the new CL Doctoral Dissertation Award at the conference. Other notable innovations include virtual presentation sessions held in parallel with the in-person poster sessions, a reception event dedicated to Findings presentations, and the introduction of combined oral and panel sessions designed to foster more thorough and interactive research discussions.

This conference could only have been made possible thanks to the invaluable work of our ACL Business Manager and Director of Events, **Jenn Rachford**, and her team — in particular, **Megan Haddad**, our ACL Administrator and Registration Manager — especially given the largest number of participants ever at ACL, estimated at the time of writing to be around 5000! Thanks also go to **Damira Mršic** and the Underline team for managing all content for the hybrid experience. The Local Organization Chairs, **Benjamin Roth** and **Dagmar Gromann**, worked hard on the bidding process, providing huge support with all local arrangements, the social event, and the visa process, eventually ensuring an outstanding experience for our in-person attendees. Our Visa Chairs — **Rexhina Blloshmi** and **Eleni Ilkou** — have helped numerous participants with their visa needs and any logistical difficulties in attending the conference. I also thank our Student Volunteer Chairs — **Pedro Henrique Luz de Araujo** and **Eleonora Mancini** — for their behind-the-scenes work, which will prove indispensable during the conference.

Additional thanks go to **Zhu Liu, Mingyang Wang**, and **Jin Zhao**, our Student Research Workshop Chairs, who, with the support of faculty advisors **Lea Frermann, Daniel Hershcovich**, and **Tristan Miller**, devoted their work to supporting and mentoring the next generation of researchers. They also secured additional funding from the Vienna Meeting Fund — congratulations!

Our Publication Chairs — **Pierpaolo Basile, Libo Qin**, and **Zhenghao Liu** — have ensured timely and high-quality conference proceedings. I also thank the Handbook Chairs, **Els Lefever and Qiongkai Xu**, for assembling the handbook that will guide attendees through the conference program.

Of course our conference could not exist without the support of our sponsors! We are deeply grateful to our Sponsorship Chairs — **Raffaella Bernardi** and **Thomas Scialom** — and to **Chris Callison-Burch**, ACL's Sponsorship Director, for securing the generous support that enables us to keep the conference accessible to as many participants as possible.

Fostering an inclusive environment remains a key goal of ACL. Our Diversity and Inclusion Chairs — **Senja Pollak, Maria Ryskina, Shane Storks**, and **Hwaran Lee** — have worked to support diverse participation and organize activities that reflect the global nature of our vibrant community. An inclusive ACL allows researchers from all areas and backgrounds to contribute, collaborate, and thrive — enriching both the scientific conversation and the broader societal impact of our work. We remain committed to building a community where everyone feels welcomed and valued.

Thank you to the Publicity and Social Media Chairs — **Anette Frank, Shruti Rijhwani**, and **Horacio Saggion** — for their communication work through social media, extending the conference's reach to a wider audience. And this brings me to our Website and App Chairs — **Xudong Han** and **Alessandro Raganato** — who kept our community informed with timely updates about all aspects of the conference. The huge number of emails and communications that we received was smoothly managed throughout the year by our Internal Communication Chairs, **Sara Tonelli** and **Yiquan Wu**. We are very grateful!

I am particularly proud of the creation of a new role, the Documentation Chair, taken by **Chenghua Lin**, who did strenuous work to assemble a new handbook with up-to-date information coming from all chairs — whom I also thank for their valuable contributions — and maintain a clear and consistent record of all our processes. Given the speed at which all conference roles evolve in their duties and commitments, I hope this role will become standard in upcoming conferences to ensure a smooth transition.

Last but not least, I wish to thank the ACL Executive Committee for their guidance, and my fellow past General and Program Chairs, and especially **Claire Gardent** and **Colin Cherry**, for generously sharing their experience and insights.

ACL 2025 is a community-wide effort. Whether you are presenting a paper, leading a workshop or tutorial, volunteering, mentoring, supporting or attending to learn and connect, your presence and participation make a difference at this conference. Thank you for being part of it!

Welcome to Vienna. Welcome to ACL 2025!

Roberto Navigli
Sapienza University of Rome & Babelscape
General Chair, ACL 2025

Program Chair Message

Welcome to the 63rd Annual Meeting of the Association for Computational Linguistics!
ACL 2025 will be held in a hybrid format, offering attendees the option to join us in person in vibrant Vienna, Austria, or to participate remotely from anywhere in the world. Organizing ACL 2025 has been a collaborative effort, made possible by the dedication and hard work of thousands of people. We gratefully acknowledge the support and contributions of the following people:

- The General Chair, Roberto Navigli;
- The ARR Editors-in-Chief of the February 2025 cycle (Jun Suzuki, Jing Jiang, and Xiaodan Zhu) and the entire team (Mausam, Viviane Moreira, Vincent Ng, Lilja Øvrelid, Anna Rogers, Michael White, Margot Mieskes, Sarvnaz Karimi);
- The techical OpenReview chairs, Niket Tandon, Lizhen Qu, and the OpenReview support team, in particular Rachel, for multiple rounds of technical help in setting up ACL 2025 on the Open Review platform;
- The 169 Senior Area Chairs;
- The 1,937 Area Chairs and the 11,720 reviewers;
- The best paper committee chairs, Rada Mihalcea and Roi Reichart, and the best paper committee members;
- The ethics chairs, Karën Fort and Bjorn Ross;
- The workshop chairs, Terra Blevins and Christophe Gravier;
- The tutorial chairs, Yuki Arase, David Jurgens and Fei Xia;
- The industry track chairs, Yunyao Li and Georg Rehm;
- The demonstration chairs, Pushkar Mishra, Smaranda Muresan, and Tao Yu;
- The internal communications chairs, Sara Tonelli and Yiquan Wu;
- The website and conference app chairs, Xudong Han and Alessandro Raganato;
- The publication chairs, Pierpaolo Basile, Libo Qin, and Zhenghao Liu;
- The handbook chairs, Els Lefever and Qiongkai Xu;
- The local organization chairs, Benjamin Roth and Dagmar Gromann, and their team;
- The visa chairs, Rexhina Blloshmi and Eleni Ilkou;
- The publicity and social media chairs, Anette Frank, Shruti Rijhwani and Horacio Saggion;
- The documentation chair, Chenghua Lin;
- The student research workshop chairs, Zhu Liu, Mingyang Wang and Jin Zhao;
- The student research workshop chairs faculty advisors, Lea Frermann, Daniel Hershcovich and Tristan Miller;

- The student volunteer chairs, Pedro Henrique Luz de Araujo and Eleonora Mancini;
- The diversity and inclusion chairs, Senja Pollak, Maria Ryskina, Shane Storks and Hwaran Lee;
- The sponsorship chairs, Raffaella Bernardi and Thomas Scialom;
- The virtual infrastructure chairs, Manling Li, Yang Liu and Avi Sil;
- The ACL Anthology Director Matt Post and his team;
- The TACL editors-in-chief (Asli Celikyilmaz, Roi Reichart, Dilek Hakkani-Tur) and CL Editor in-Chief Wei Lu for coordinating TACL and CL presentations with us;
- The ACL 2024 Program Chairs, Lun-Wei Ku, André F. T. Martins, Vivek Srikumar, for information and support;
- Damira Mrsic and Underline Team;
- Jennifer Rachford and entire conference support staff;
- All the authors of papers who submitted their papers for review in the ARR 2025 February cycle and those who committed to the ACL 2025 conference.

Review Process All submissions to ACL 2025 went through a two-stage review process. First, papers were submitted to the ACL Rolling Review (ARR), where they were reviewed by reviewers and received meta-reviews from Area Chairs. Then, authors had the option to commit their reviewed papers to ACL via a separate ACL 2025 commitment site. At this stage, Senior Area Chairs provided recommendations, and final acceptance decisions were made by the Program Chairs. This process is consistent with previous conferences, including ACL 2024, EACL 2024, and NAACL 2025.
We worked closely with the ARR team, particularly the ARR February 2025 Editors-in-Chief, and served as guest Editors-in-Chief for this round. We helped recruit new reviewers and Area Chairs to ARR, resulting in 11,720 reviewers and 1,942 Area Chairs in the February 2025 ARR cycle to which most ACL 2025 papers were submitted. ACL also recruited 169 Senior Area Chairs to oversee the review and meta-review process. Overall, the ARR process ran smoothly, ensuring that all submitted papers received at least three reviews and a meta-review. For the ACL commitment phase, Senior Area Chairs made recommendations for 5,356 committed papers based on the reviews, meta-reviews, and the papers themselves, with final acceptance decisions made by the Program Chairs.

Acceptance Rate
In total, there are 1,699 papers accepted to the Main Conference and 1,392 papers accepted to Findings. The acceptance rate calculation follows precedent set by previous conferences that go through ACL Rolling Review (ARR), e.g. ACL 2024. The calculation takes into account the multi-stage process of ARR where a paper may get revised in ARR and then later committed to the conference. In total, we had 8,360 unique submissions across the December 2024 and February 2025 ARR cycles of which 5,501 papers were committed to ACL. The acceptance rate is 20.3% for the Main Conference papers and a further 16.7% for Findings papers.

Special Theme: Generalization of NLP Models
ACL 2025's special theme is generalization of NLP models. Generalization is crucial for ensuring that models behave robustly, reliably, and fairly when making predictions on data different from their training data. Achieving good generalization is critically important for models used in real-world applications, as they should emulate human-like behavior. Humans are known for their ability to generalize well, and models should aspire to this standard. The theme track invites empirical and theoretical research and position and survey papers reflecting on the Generalization of NLP Models. The possible topics of discussion

include (but are not limited to) the following: 1. How can we enhance the generalization of NLP models across various dimensions—compositional, structural, cross-task, cross-lingual, cross-domain, and robustness? 2. What factors affect the generalization of NLP models? 3. What are the most effective methods for evaluating the generalization capabilities of NLP models? 4. While Large Language Models (LLMs) significantly enhance the generalization of NLP models, what are the key limitations of LLMs in this regard?

We received 128 submissions to the theme track during the review phase. Among these, 41 papers were accepted to the main conference and a further 33 to Findings of ACL 2025. The conference will also feature a panel discussion on the theme of Generalization, with the participation of leading experts in this area.

Best Paper Selection

ACL 2025 implemented the updated ACL award policy that seeks to expand the pool of work recognized as outstanding. In total 117 papers were nominated by the reviewers, area chairs and senior area chairs for best paper consideration. The best paper committee assessed these papers to select the best papers (featuring $(\leq)0.6\%$ of accepted papers), outstanding papers (featuring $(\leq)2.5\%$ of accepted papers), and special awards for social impact and best resource. Based on the review by the Best Paper committee, 39 papers have been selected for awards in the above categories. Separately, the senior area chairs also nominated their favourite papers as SAC Highlights.

In addition, we have awards for test of time award for a paper published in TACL in 2013 or 2014 and the best paper award for a paper published in TACL in 2024. The final selection was made by the best paper committee, and the winners will be announced during the closing ceremony. The ACL 2025 Best Papers will also be given an opportunity to present their work in the closing ceremony.

Program Composition & Presentation Modes

Based on feedback from the conference support staff and the Underline team after ACL 2025, we decided to hold the virtual presentations sessions during the main conference. This enables us to align the virtual sessions with time slots when in-person participants are available. This approach allows virtual attendees to participate concurrently with the physical event, avoiding the need for organizers and attendees to engage with the conference twice and separately. This year, 218 main conference papers were selected for oral presentations by the program chairs, with the goal of creating a well-rounded program featuring a diverse set of topics instead of selecting papers based on their review scores. This year we have introduced a panel section of the conference. Out of the selected oral presenters, 25 will have an opportunity to not only present their work but also participate in a panel discussion. We have lined up five panels on the theme of: Generalisation in NLP, LLM alignment, Human-centred NLP, Interpretability and model analysis, Multilinguality and language diversity. In addition to the main conference papers, the ACL program also includes 18 papers accepted by Computational Linguistics and 42 papers accepted by Transactions of the ACL (TACL). Among these, 11 journal papers will be presented in-person as oral presentations, thematically distributed across appropriate sessions.

This year the conference will also for the first time feature a dedicated Findings poster session with reception. All Findings presenters have been assigned poster presentations in this session or alongside other posters for main conference papers in the same track. Rounding out the program are dedicated sessions for the demonstrations track and the student research workshop.

Keynotes and Panel

This year's program features an impressive lineup of two keynote presentations:

- Prof. Luke Zettlemoyer from Paul G. Allen School of Computer Science Engineering at the University of Washington, and a Senior Research Director at Meta, will share his insights on "Rethinking Pretraining: Data and Architecture."

- Prof. Verena Rieser from Google DeepMind will present on "Whose Gold? Re-imagining Alignment for Truly Beneficial AI."

Alongside these keynotes, we are thrilled to host a panel discussion aiming to answer the question Can large language models (LLMs) generalize?. Our esteemed panelists include:

- Prof. Eduard Hovy, University of Melbourne (who will also act as the panel chair).
- Prof. Mirella Lapata, University of Edinburgh
- Prof. Yue Zhang, Westlake University
- Prof. Dan Roth, University of Pennsylvania and Oracle

This diverse group of panelists will provide a comprehensive view of the latest trends and challenges in Generalisation of NLP in the Large Language Models era.

We hope you enjoy this year's diverse and engaging program!

Ekaterina Shutova (University of Amsterdam and Stanford University)
Mohammad Taher Pilehvar (Cardiff University and Tehran Institute for Advanced Studies)
Joyce Nakatumba-Nabende (Makerere University)
Wanxiang Che (Harbin Institute of Technology)

ACL 2025 Program Co-Chairs

ACL Business and Event Manager Message

On behalf of the ACL Executive team, I would like to extend my sincere gratitude to everyone who contributed to making ACL 2025 a reality.
A special thank you to our Local Committee—**Benjamin Roth** and **Dagmar Gromann**—as well as **Markus Heigl** from the Vienna Convention Bureau, our destination partner Meeting Destination Vienna, for their exceptional work throughout the submission and selection process. Your detailed planning and unwavering support helped bring ACL to this beautiful and historic city.
I also wish to thank the **ACL Boards** for their thoughtful review of venue bids and their decision to select Vienna, Austria as the host city. A particular note of appreciation goes to **David Yarowsky** for his guidance and efforts in finalizing our venue contracts. To our organizational partners and extended team:

- **ACL Onsite Team**
  Megan Haddad – Event Assistant & ACL Admin
- **Auxiliary Staff**
  Lina Staggs, Sally Stevenson, Brandy Dorsey, Karen Maloy, Jason Staggs
- **Support Staff – Lee Hartman & Sons**
  Jon Dorsey – AV Director
  Trevor Laffoon – AV Tech
- **Underline – Hybrid Event Partner**
  Damira Mrsic – Content & Hybrid Manager
- **On-Site Support Team**
  Jernej Masnec, Borna Bevanda, Leticija Maričić, Petra Filipčić, Petra Vižintin, Rafael Grabovica, Lucijan Prpic, Nikola Pesut, Astrid Plišo, Kristijan Varga, Filip Cigic, Mihael Herenda, Petar Vođinac, Roko Mirković
- **Austria Center Vienna**
  Friedrich Gehmacher – for your guidance and help in contracting the venue
  Victoria Wozniak – for managing the on-site logistics with care and precision
- **Vendor Partners**
  Motto Catering
  Stand Out (Exhibition Services)
  XEST(Print Services)
  Marcelo Viridian (Graphics)
- **Hotel Partners** Meliá Vienna
  NH Danube City
  Spark by Hilton Vienna Donaustadt

Finally, I want to express my heartfelt thanks to the entire **organizing and program committees**. Your dedication, long hours, and often sleepless nights have made ACL 2025 possible. Your commitment to academic excellence and community engagement is what makes this event a continued success year after year.

I look forward to a memorable and enriching conference here in Vienna, and I thank each of you for being a part of it. With gratitude,

Jennifer Rachford
Business and Event Manager

Organizing Committee

General Chair

Roberto Navigli, Sapienza University of Rome & Babelscape

Program Chairs

Wanxiang Che, Harbin Institute of Technology
Joyce Nakatumba-Nabende, Makerere University
Ekaterina Shutova, University of Amsterdam
Mohammad Taher Pilehvar, Cardiff University / Tehran Institute for Advanced Studies

Local Chairs

Benjamin Roth, University of Vienna
Dagmar Gromann, University of Vienna

Workshop Chairs

Terra Blevins, University of Vienna
Christophe Gravier, Université Jean Monnet

Tutorial Chairs

Yuki Arase, Tokyo Institute of Technology
David Jurgens, University of Michigan
Fei Xia, University of Washington

Demonstration Chairs

Pushkar Mishra, Google DeepMind / Cambridge
Smaranda Muresan, Barnard College, Columbia University
Tao Yu, The University of Hong Kong

Student Research Workshop Chairs

Zhu Liu, Tsinghua University
Mingyang Wang, LMU
Jin Zhao, Brandeis University

Student Research Workshop Chair: Faculty Advisors

Lea Frermann, University of Melbourne
Daniel Hershcovich, University of Copenhagen
Tristan Miller, University of Manitoba

Publication Chairs

Pierpaolo Basile, University of Bari
Libo Qin, Central South University
Zhenghao Liu, Northeastern University

Handbook Chairs

Els Lefever, Ghent University
Qiongkai Xu, Macquarie University

Sponsorship Chairs

Raffaella Bernardi, University of Trento
Thomas Scialom, Meta AI

Diversity and Inclusion Chairs

Senja Pollak, Jožef Stefan Institute
Maria Ryskina, MIT
Shane Storks, University of Michigan
Hwaran Lee, NAVER

Publicity and Social Media Chairs

Anette Frank, Heidelberg University
Shruti Rijhwani, Google DeepMind
Horacio Saggion, Pompeu Fabra University

Website and Conference App Chairs

Xudong Han, LibrAI / MBZUAI
Alessandro Raganato, University of Milano-Bicocca

Ethics Chairs

Karën Fort, Université de Lorraine / LORIA
Bjorn Ross, University of Edinburgh

Student Volunteer Chairs

Pedro Henrique Luz de Araujo, University of Vienna
Eleonora Mancini, University of Bologna

Technical Open Review Chairs

Niket Tandon, Allen AI
Lizhen Qu, Monash University

Virtual Infrastructure Chairs

Manling Li, Northwestern University
Yang Liu, Microsoft
Avi Sil, IBM Research

Internal Communication Chairs

Sara Tonelli, FBK

Yiquan Wu, Zhejiang University

Industry Track Chairs

Yunyao Li, Adobe
Georg Rehm, DFKI

Best Paper Committee Chairs

Rada Mihalcea, University of Michighan
Roi Reichart, Technion - Israel Institute of Technology

Visa Chairs

Rexhina Blloshmi, Amazon AGI
Eleni Ilkou, TIB

Documentation Chair

Chenghua Lin, University of Manchester

ARR Editors in Chief

Jun Suzuki, Tohoku University
Jing Jiang, Australian National University
Xiaodan Zhu, Queen's University

Program Committee

Senior Area Chairs

David Ifeoluwa Adelani, McGill University
Alan Akbik, Humboldt Universität Berlin
Raquel G. Alhama, University of Amsterdam, University of Amsterdam
Afra Alishahi, Tilburg University
Jacob Andreas, Massachusetts Institute of Technology
Maria Antoniak
Marianna Apidianaki, University of Pennsylvania, University of Pennsylvania
Jun Araki, Bosch Research
Ehsaneddin Asgari, Qatar Computing Research Institute and University of California, Berkeley
Edoardo Barba, University of Roma "La Sapienza"
Jasmijn Bastings, Google DeepMind
Beata Beigman Klebanov, Educational Testing Service
Ahmad Beirami
Yonatan Belinkov, Technion, Technion
Meriem Beloucif, Uppsala University
Jonathan Berant, Google and Tel Aviv University
Steven Bethard, University of Arizona
Alexandra Birch, University of Edinburgh
Jelke Bloem, University of Amsterdam
Danushka Bollegala, Amazon and University of Liverpool
Houda Bouamor, Carnegie Mellon University
Jordan Lee Boyd-Graber, University of Maryland, College Park
Emanuele Bugliarello, Google DeepMind
Jan Buys, University of Cape Town
Aoife Cahill, Dataminr
Jose Camacho-Collados, Cardiff University
Hanjie Chen, Rice University
Mingda Chen, Meta FAIR
Danqi Chen, Princeton University
Christos Christodoulopoulos, Amazon
Arman Cohan, Yale University and Allen Institute for Artificial Intelligence
Simone Conia, Sapienza University of Rome
Ryan Cotterell, Swiss Federal Institute of Technology
Yiming Cui, iFLYTEK Research
Raj Dabre, Department of Computer Science, Indian Institute of Technology, Madras, Indian Institute of Technology, Madras and National Institute of Information and Communications Technology (NICT), National Institute of Advanced Industrial Science and Technology
Tejaswini Deoskar, Utrecht University
Leon Derczynski, NVIDIA and IT University of Copenhagen
Yuxiao Dong, Tsinghua University
Li Dong, Microsoft Research
Yue Dong, University of California, Riverside and McGill University
A. Seza Doğruöz, Ghent University
Greg Durrett, University of Texas at Austin
Ondrej Dusek, Charles University, Prague
Nouha Dziri
Desmond Elliott, Copenhagen University and University of Copenhagen
Guy Emerson, University of Cambridge
Alexander Fabbri, Scale AI

Anna Feldman, Montclair State University and Montclair State University
Fuli Feng, University of Science and Technology of China
Xiaocheng Feng
Anjalie Field, Johns Hopkins University
Michael Flor, Educational Testing Service
Antske Fokkens, VU University Amsterdam
Zhe Gan, Apple
Albert Gatt, Utrecht University
Mor Geva, Tel Aviv University and Google Research
Marjan Ghazvininejad, Facebook
Goran Glavaš, Julius-Maximilians-Universität Würzburg
Yvette Graham
Jiafeng Guo, Institute of Computing Technolgy, Chinese Academy of Sciences
Xianpei Han, Institute of Software, CAS
Peter Hase, Schmidt Sciences
Devamanyu Hazarika, Facebook
Yulan He, King's College London, University of London
Mohammad Javad Hosseini, Google
Neil Houlsby, Google
Dieuwke Hupkes, Facebook
Ignacio Iacobacci, Elm Europe
Mohit Iyyer, University of Maryland, College Park
Zhijing Jin, Department of Computer Science, University of Toronto
Shafiq Joty, Nanyang Technological University and SalesForce.com
Rabeeh Karimi Mahabadi, NVIDIA
Frank Keller, University of Edinburgh
Daniel Khashabi, Johns Hopkins University
Najoung Kim, Boston University
Svetlana Kiritchenko, National Research Council Canada
Ekaterina Kochmar, Mohamed bin Zayed University of Artificial Intelligence
Julia Kreutzer, Cohere for AI
Kalpesh Krishna, Google DeepMind
Ranjay Krishna, University of Washington
Matthieu Labeau, Télécom ParisTech
Anne Lauscher, Universität Hamburg
Carolin Lawrence, NEC Laboratories Europe and NEC Laboratories Europe
Wenqiang Lei, Sichuan University
Martha Lewis, University of Amsterdam
Junyi Jessy Li, University of Texas at Austin
Zhenghua Li, Soochow University
Dongxu Li, SalesForce.com
Piji Li, Nanjing University of Aeronautics and Astronautics
Bill Yuchen Lin, xAI and University of Washington
Yiqun Liu, Tsinghua University
Fangyu Liu, Google DeepMind
Zhiyuan Liu, Tsinghua University
Kang Liu, Institute of automation, Chinese academy of science, Chinese Academy of Sciences
Edison Marrese-Taylor, The Univesity of Tokyo and AIST, National Institute of Advanced Industrial Science and Technology
Joshua Maynez, Google
Dipendra Misra, Mosaic Research, Databricks
Saif M. Mohammad
Nafise Sadat Moosavi, University of Sheffield

Aaron Mueller, Northeastern University and Technion - Israel Institute of Technology
Philippe Muller, IRIT, University of Toulouse
Preslav Nakov, Mohamed bin Zayed University of Artificial Intelligence
Vivi Nastase, University of Geneva
Wolfgang Nejdl, L3S Research Center and Universität Hannover
Xuefei Ning, Tsinghua University
Debora Nozza, Bocconi University
Giorgio Di Nunzio, University of Padua
Nedjma Ousidhoum, Cardiff University
Alexandros Papangelis, Apple
Gabriella Pasi, University of Milano-Bicocca
Rebecca J. Passonneau, Pennsylvania State University
Viviana Patti, University of Turin
Hao Peng, University of Illinois Urbana-Champaign
Carla Perez-Almendros, Cardiff University
Tiago Pimentel, Department of Computer Science, ETHZ - ETH Zurich
Edoardo Ponti, University of Edinburgh
Simone Paolo Ponzetto, Universität Mannheim
Daniel Preotiuc-Pietro, Bloomberg
Emily Prud'hommeaux, Boston College
Peng Qi, Orby AI
Marek Rei, Imperial College London
Verena Rieser, Google DeepMind
Roland Roller, German Research Center for AI
Alla Rozovskaya
Michael Sejr Schlichtkrull, Queen Mary, University of London
Steven Schockaert, Cardiff University
Rico Sennrich, University of Zurich
Ehsan Shareghi, Monash University
Qinlan Shen, Oracle
Kevin Small, Amazon
Mihai Surdeanu, University of Arizona
Anders Søgaard, Copenhagen University
Chenhao Tan, University of Chicago
Hao Tan, Adobe Systems
Harish Tayyar Madabushi, University of Bath
Alberto Testoni, Amsterdam UMC
Trang Tran, N/A
Marcos Vinicius Treviso
Yulia Tsvetkov, Department of Computer Science, University of Washington
Daniel Varab, German Research Center for AI
Suzan Verberne, Universiteit Leiden
Karin Verspoor, Royal Melbourne Institute of Technology
Ivan Vulić, Google DeepMind and University of Cambridge
Henning Wachsmuth, Leibniz Universität Hannover
Byron C Wallace, Northeastern University, Brown University and Northeastern University
Qifan Wang, Meta AI
Tong Xiao, Northeastern University
Deyi Xiong, Tianjin University
Ruifeng Xu, Harbin Institute of Technology
Yadollah Yaghoobzadeh
Diyi Yang, Stanford University
Min Yang, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Chinese

Academy of Sciences
Helen Yannakoudakis, King's College London, University of London
Huaxiu Yao, Department of Computer Science, University of North Carolina at Chapel Hill
Andrew Yates, Johns Hopkins University
Kai Yu, Shanghai Jiao Tong University
Luke Zettlemoyer, University of Washington, Facebook and Meta
Weinan Zhang, Harbin Institute of Technology
Meishan Zhang, Harbin Institute of Technology (Shenzhen), China
Yue Zhang, Westlake University
Qi Zhang, Fudan University
Xin Zhao, Renmin University of China
Yi Zhou, Cardiff University
Arkaitz Zubiaga, Queen Mary University of London

2

Anti-harassment Policy

ACL 2025 adheres to the ACL Anti-Harassment Policy. Any participant who experiences harassment or hostile behavior may contact any current member of the ACL Professional Conduct Committee or Jennifer Rachford, who is usually available at the registration desk of the conference. Please be assured that if you approach us, your concerns will be kept in strict confidence, and we will consult with you on any actions taken. The open exchange of ideas, the freedom of thought and expression, and respectful scientific debate are central to the aims and goals of a ACL conference. These require a community and an environment that recognizes the inherent worth of every person and group, that fosters dignity, understanding, and mutual respect, and that embraces diversity. For these reasons, ACL is dedicated to providing a harassment-free experience for participants at our events and in our programs. Harassment and hostile behavior are unwelcome at any ACL conference. This includes speech or behavior (including in public presentations and on-line discourse) that intimidates, creates discomfort, or interferes with a persons participation or opportunity for participation in the conference. We aim for ACL conferences to be an environment where harassment in any form does not happen, including but not limited to: harassment based on race, gender, religion, age, color, national origin, ancestry, disability, sexual orientation, or gender identity. Harassment includes degrading verbal comments, deliberate intimidation, stalking, harassing photography or recording, inappropriate physical contact, and unwelcome sexual attention. The ACL board members are listed at: https://www.aclweb.org/portal/about. The full policy and its implementation is defined at: https://aclweb.org/adminwiki/index.php/Anti-Harassment_Policy

Ethics Policy

ACL adopts the ACM Code of Ethics (https://www.acm.org/code-of-ethics) in the version adopted June 22nd, 2018, by the ACM Council. In its application to ACL, it is to be read in the contextually appropriate interpretation, e.g., ACM member is to be read as ACL member. Sec 4.2 should be read as follows: 4.2 Treat violations of the Code as inconsistent with membership in the ACL. Each ACL member should encourage and support adherence by all members of the CL/NLP community regardless of ACL membership. ACL members who recognize a breach of the Code should consider reporting the violation to the ACL, which may result in remedial action. The open exchange of ideas, the freedom of thought and expression, and respectful scientific debate are central to the aims and goals of the ACL. These require a community and an environment that recognizes the inherent worth of every person and group, that fosters

dignity, understanding, and mutual respect, and embraces diversity. For these reasons, ACL is dedicated to providing a harassment-free experience for all the members, as well as participants at our events and in our programs. Harassment and hostile behavior are unwelcome at any ACL conference, associated event, or in ACLaffiliated online discussions. This includes speech or behavior that intimidates, creates discomfort, or interferes with a persons participation or opportunity for participation in a conference or an event. We aim for ACL-related activities to be an environment where harassment in any form does not happen, including but not limited to: harassment based on race, gender, religion, age, color, appearance, national origin, ancestry, disability, sexual orientation, or gender identity. Harassment includes degrading verbal comments, deliberate intimidation, stalking, harassing photography or recording, inappropriate physical contact, and unwelcome sexual attention. The policy is not intended to inhibit challenging scientific debate, but rather to promote it by ensuring that all are welcome to participate in the shared spirit of scientific inquiry. Vexatious complaints and willful misuse of this procedure will render the complainant subject to the same sanctions as a violation of the anti-harassment policy. It is the responsibility of the community as a whole to promote an inclusive and positive environment for our scholarly activities. In addition, anyone who experiences harassment or hostile behavior may contact any member of the Professional Conduct Committee. Members of this committee are instructed to keep any such contact in strict confidence, and those who approach the committee will be consulted before any actions are taken.

3 Meal Info

Breakfast

July 27 - August 1
Breakfast is not provided, you can check with your local hotel and near by restaurants.

Breaks †

July 27 - August 1
Coffee, tea, pastry, and fruit are provided late morning and midafternoon.

Lunch

July 27 - August 1
Lunch is not provided, however, the Austria Center Vienna will have lunch options available for purchase. See detailed agenda for times.

Dinner

July 27 - August 1
Dinner is not provided, but there are plenty of cafes, and restaurants within the area.

Social Events ††

Welcome Reception, July 27, 18:30 - 20:30

Light canapes, and a drink ticket will be provided on Sunday, July 27, 2025, at the Welcome Reception. It will be held on Level 0 Hall E of the Austria Center Vienna.

Findings Reception, Monday, July 28, 2025 †††

Light canapes, and refreshments will be provided on Monday, July 28, 2025, at the Findings Recpetion. It will be held on Level 0 Hall E of the Austria Center Vienna.

Social Event, July 29, 19:00 - 22:00

Social Event Gala Dinner International Buffet Dinner and a drink ticket will be provided on Tuesday, July 29, 2025, at the Social Event Gala Dinner (Social Gala). Located in Level 2 Hall A of the Austria Center Vienna.

†: Included in Main Conference, Tutorial, and Workshop registrations.
††: Included in the Main Conference registration only, and can be added on for Guests, Tutorials, Workshops, and Exhibitors to attend at the Registration Solutions Desk or through your Events registration login. **No admission without an entry ticket**
†††: Included in the Main Conference registration only. No guest add ons for this reception



Skip the line & preorder your warm lunch

We look forward to welcoming all participants to the Annual Meeting of the Association for Computational Linguistics 2025 in Vienna and to our **various gastronomy outlets** at the Austria Center Vienna.

Here is your overview of the gatronomy outlets:

- Café Motto, indoor, level 0
- Foyer A, indoor, level 2
- Market stand, indoor, level 1
- Market stand, indoor, hall X5
- Snack hut with Viennese sausage stand, outdoor, front of the main entrance
- Lunch market & beergarden, outdoor, forecourt of the building
- Coffee stations, outdoor, front of the main entrance

In addition to buying food, snacks and drinks on site - you have the option of pre-ordering your warm lunch.

Warm lunch will be served from Sunday, 27th July to Friday, 1st August 2025 from 11am to 2pm at the forecourt of the building. You are already invited to pre-order your lunch for Sunday, 27th July 2025. You can then order your lunch for the following day every day from 2:00 pm directly in our webshop.

Make your order and we'll meet at the fast lane!

MOTTO CATERING

Your MOTTO-CATERING Team.

LINK

https://my.apro.at/web/locations/10886?menu=930499

4 Social Events

ACL 2025 Welcome Reception

Venue: Austria Center Vienna, Level 0 Hall E
Time: Sunday, July 27, 2025 18:30 - 20:30

Kick off **ACL 2025** with an evening of networking, refreshments, and engaging conversations at our **Welcome Reception**! This gathering is the perfect opportunity to connect with fellow attendees, reconnect with colleagues, and immerse yourself in the vibrant natural language processing community before the conference begins in full.
Enjoy a relaxed atmosphere with light hors d'oeuvres and drinks as you meet researchers, practitioners, and industry leaders from around the world. Whether you're a first-time attendee or a returning participant, this reception sets the stage for an inspiring and productive conference experience.

Important Information:
Attendance is included for all Main Conference registrants. If you are not registered for the Main Conference but would like to attend, you may add this event to your registration at the Registration Solution Desk located on the lobby level.
We look forward to welcoming you to **ACL 2025**!

Warm regards,
The ACL 2025 Organizing Committee

ACL 2025 Findings Reception

Venue: Austria Center Vienna, Level 0 Hall X4 X5
Time: Monday, July 28, 2025 18:00 - 19:00

Join us for an engaging evening as we kick off the evening with a relaxed reception featuring light snacks and refreshments. Mingle with fellow attendees, enjoy a drink, and explore the poster presentations showcasing innovative findings. It's the perfect setting to connect, collaborate, and spark ideas.

Important Information:
Attendance is included for all Main Conference registrants only.

Warm regards,
The ACL 2025 Organizing Committee

ACL 2025 Social Event

Venue: Austria Center Vienna, Level 2 Hall A
Time: Tuesday, July 29, 2025 19:00 - 22:00

Join us for an unforgettable evening celebrating the spirit of Vienna, Austria. A perfect fusion of refinement, cultural exploration, and high-spirited fun, this event promises unforgettable moments for all in attendance.

Entertainment Lineup: Enjoy the graceful artistry of Waltz Ballet. An exciting introduction to Voigasplattler, an energetic folk dance that showcases rhythmic precision and vibrant footwork. An electrifying performance by the Sugar Swing Cats.
Enjoy a delicious buffet dinner featuring regional flavors. **Beer & wine tickets** available for guests.
This is your chance to unwind, meet fellow attendees, and experience the best of Vienna's artistic charm. We look forward to celebrating with you!

Important Information: **Attendance is included** for all Main Conference registrants. If you are not registered for the Main Conference but would like to attend, you may **add this event to your registration** at the Registration Solutions Desk.

Warm regards,
The ACL 2025 Organizing Committee

5 Keynotes

Monday, July 28

Rethinking Pretraining: Data and Architecture



Luke Zettlemoyer
Monday, July 28, 09:30 - 10:30

Abstract: Large language model training follows a standard pipeline: tokenization, pretraining, possibly mid-training, and post training or alignment. Despite its wild success, we understand relatively little about this recipe and are almost certainly missing many opportunities to improve it. In this talk, I will focus on three such cases. I'll describe our work on data efficient post training (e.g. LIMA, ALMA, and s1) where we argue that nearly all advanced model capabilities ultimately come from the pretraining data, even if effective alignment is still essential for controlling model behavior. I will also describe new methods for extracting more signal from the pretraining data, including new hierarchical architectures for byte-level language models (e.g. BLT) that are both tokenizer-free and scale better than traditional BPE-based methods, especially in the long tail. Finally, I will discuss decentralized, modular training algorithms

(e.g. BTM) that better isolate and control the influence of specific data on specific model components and behaviors. Together, these methods promise to simplify training and improve scaling, by centering and amplifying the influence of data in architecture design.

Bio: Luke Zettlemoyer is a Professor in the Paul G. Allen School of Computer Science & Engineering at the University of Washington, and a Senior Research Director at Meta. His research focuses on empirical methods for natural language semantics, and involves designing machine learning algorithms, introducing new tasks and datasets, and, most recently, studying how to best develop new architectures and self-supervision signals for pre-training. His honors include being elected ACL President, named an ACL Fellow, winning a PECASE award, an Allen Distinguished Investigator award, and multiple best paper awards. Luke was an undergrad at NC State, received his PhD from MIT and was a postdoc at the University of Edinburgh.

Tuesday, July 29

Whose Gold? Re-imagining Alignment for Truly Beneficial AI



Verena Rieser

Tuesday, July 29, 09:00 - 10:00

Abstract: Human feedback is often the "gold standard" for AI alignment, but what if this "gold" reflects diverse, even contradictory human values? This keynote explores the technical and ethical challenges of building beneficial AI when values conflict – not just between individuals, but also within them. My talk advocates for a dual expansion of the AI alignment framework: moving beyond a single, monolithic viewpoint to a plurality of perspectives, and transcending narrow safety and engagement metrics to promote comprehensive human well-being.

Bio: Verena Rieser is a Senior Staff Research Scientist at Google DeepMind, where she founded the VOICES team (Voices-of-all in alignment). Her team is a core contributor to Gemini with a mission to enhance model safety and usability for diverse communities. Verena has pioneered work in data-driven multimodal Dialogue Systems and Natural Language Generation, encompassing conversational RL agents, faithful data-to-text generation, spoken language understanding, evaluation methodologies, and applications of AI for societal good. Verena previously directed the NLP lab as a full professor at Heriot-Watt University, Edinburgh, and held a Royal Society Leverhulme Senior Research Fellowship. She earned her PhD from Saarland University.

6 Panel Discussion

Monday, July 28

Panel Discussion: Generalization of NLP Models






Panel: Mirella Lapata, Dan Roth, Yue Zhang
Moderator: Eduard Hovy
Monday, July 28, 16:30 - 17:30

Mirella Lapata is professor of natural language processing in the School of Informatics at the University of Edinburgh. Her research focuses on getting computers to understand, reason with, and generate natural language. She is the first recipient (2009) of the British Computer Society and Information Retrieval Specialist Group (BCS/IRSG) Karen Sparck Jones award and a Fellow of the Royal Society of Edinburgh, the ACL, and Academia Europaea. Mirella has also received best paper awards in leading NLP conferences and has served on the editorial boards of the Journal of Artificial Intelligence Research, the Transactions of the ACL, and Computational Linguistics. She was president of SIGDAT (the group that organizes EMNLP) in 2018. She has been awarded an ERC consolidator grant, a Royal Society Wolfson Research Merit Award, and a UKRI Turing AI World-Leading Researcher Fellowship.

Dan Roth is the Eduardo D. Glandt Distinguished Professor at the University of Pennsylvania and Chief AI Scientist at Oracle. Until June 2024 Dan was a VP/Distinguished Scientist at AWS AI where he led the scientific effort behind Amazon's first-generation GenAI products, including Titan Models, Amazon Q, and Amazon Bedrock. Dan is a Fellow of the AAAS, ACM, AAAI, and ACL, and a recipient of the IJCAI John McCarthy Award "for major conceptual and theoretical advances in the modeling of natural language understanding, machine learning, and reasoning." He has published broadly in natural language processing, machine learning, knowledge representation and reasoning, and learning theory, was the Editor-in-Chief of the Journal of Artificial Intelligence Research (JAIR) and has served as a Program Chair and Conference Chair for the major conferences in his research areas. Roth has been involved in several ML/NLP/GenAI startups in domains that range from legal and compliance to health care. Dan received his B.A Summa cum laude in Mathematics from the Technion, Israel and his Ph.D. in Computer Science from Harvard University in 1995.

Yue Zhang is a tenured Professor at Westlake University (https://frcchang.github.io). His research interests include fundamental NLP and its machine learning algorithms, and his recent research focuses on LLM reasoning and AI scientist. His major contributions to the field include machine learning algorithms for structured prediction (e.g., parsing and IE), neural NLP models (i.e., lattice and graph LSTM), and generalization for NLP/LM (e.g., OOD and logical reasoning). He co-authored the Cambridge University Press book "Natural Language Processing – a Machine Learning Perspective'' and served as a PC co-chair for CCL 2020 and EMNLP 2022, test-of-time award committee co-chairs for ACL 2024 and 2025, action editor for TACL, and associate editor for TASLP, TALLIP, TBD, and CSL. He won the best paper awards of IALP 2017 and COLING 2018, best paper honorable mention of SemEval 2020, and best paper nomination for ACL 2018 and ACL 2023.

Eduard Hovy is the Executive Director of Melbourne Connect (a research and tech transfer centre at the University of Melbourne), a professor at the University of Melbourne's School of Computing and Information Systems, and an adjunct professor at the Language Technologies Institute in the School of Computer Science at Carnegie Mellon University. In 2020–21 he served as Program Manager in DARPA's Information Innovation Office (I2O), where he managed programs in Natural Language Technology and Data Analytics. Dr. Hovy completed a Ph.D. in Computer Science (Artificial Intelligence) at Yale University in 1987 and was awarded honorary doctorates from the National Distance Education University (UNED) in Madrid in 2013 and the University of Antwerp in 2015. He is one of the initial 17 Fellows of the Association for Computational Linguistics (ACL) and is also a Fellow of the Association for the Advancement of Artificial Intelligence (AAAI). Dr. Hovy's research focuses on computational semantics of language and addresses various areas in Natural Language Processing and Data Analytics, including in-depth machine reading of text, information extraction, automated text summarization, question answering, the semi-automated construction of large lexicons and ontologies, and machine translation. In early 2025 his Google h-index was 109, with about 68,000 citations. Dr. Hovy is the author or co-editor of eight books and over 400 technical articles and is a popular invited speaker. He regularly co-taught Ph.D.-level courses and has served on Advisory and Review Boards for both research institutes and funding organizations in Germany, Italy, Netherlands, Ireland, Singapore, and the USA. From 2003 to 2015 he was co-Director of Research for the Department of Homeland Security's Center of Excellence for Command, Control, and Interoperability Data Analytics, a distributed cooperation of 17 universities. In 2001 Dr. Hovy served as

President of the international Association of Computational Linguistics (ACL), in 2001–03 as President of the International Association of Machine Translation (IAMT), and in 2010–11 as President of the Digital Government Society (DGS).

7

Tutorials

Overview

Sunday, July 27, 2025

07:30 - 20:30	***Registration***	
09:00 - 12:30	***Morning Tutorials***	
	Tutorial 1: Inverse Reinforcement Learning Meets Large Language Model Alignment Mihaela van der Schaar and Hao Sun	*Level 2 Hall B*
	Tutorial 2: Eyetracking and NLP David Reich, Omer Shubi, Lena Jäger and Yevgeni Berzak	*Level 1 Hall M*
	Tutorial 3: Uncertainty Quantification for Large Language Models Artem Shelmanov, Maxim Panov, Ekaterina Sergeevna Fadeeva, Artem Vazhentsev, Roman Konstantinovich Vashurin and Timothy Baldwin	*Level 2 Hall C*
	Tutorial 4: Human-AI Collaboration: How AIs Augment Human Teammates Tongshuang Wu, Diyi Yang, Kyle Lo and Marti A. Hearst	*Level 1 Hall N*
14:00 - 17:30	***Afternoon Tutorials***	
	Tutorial 5: Navigating Ethical Challenges in NLP: Hands-on strategies for students and researchers Luciana Benotti, Fanny Ducel, Karen Fort, Guido Ivetta, Zhijing Jin, Min-Yen Kan, Seunghun Lee, Margot Mieskes, Minzhi Li, and Adriana Pagano	*Level 1 Hall M*
	Tutorial 6: NLP for Counterspeech against Hate and Misinformation Daniel Russo, Helena Bonaldi, Yi-Ling Chung, Gavin Abercrombie and Marco Guerini	*Level 1 Hall N*

	Tutorial 7: Synthetic Data in the Era of Large Language Models Vijay Viswanathan, Xiang Yue, Alisa Liu, Yizhong Wang and Graham Neubig	*Level 2 Hall B*
	Tutorial 8: Guardrails and Security for LLMs: Safe, Secure, and Controllable Steering of LLM Applications Traian Rebedea, Leon Derczynski, Shaona Ghosh, Makesh Narsimhan Sreedhar, Faeze Brahman, Liwei Jiang, Bo Li, Yulia Tsvetkov, Christopher Parisien and Yejin Choi	*Level 2 Hall C*
18:30 - 20:30	***Welcome reception***	

Turorial 1

T1: Inverse Reinforcement Learning Meets Large Language Model Alignment

Mihaela van der Schaar and Hao Sun

Large Language Model (LLM) alignment remains one of the most critical challenges in reinforcement learning. As the success of models like DeepSeek-R1 demonstrates, improving alignment requires better architectures and a deeper understanding of reinforcement learning (RL) and reward modeling. This tutorial explores the connection between Inverse Reinforcement Learning (IRL) and LLM alignment, offering a structured roadmap for researchers and practitioners.

We frame LLM alignment as an inverse RL problem, contrasting traditional reinforcement learning with inverse methods that infer rewards from human data. A key focus is on reward models, examining how they are constructed from various data sources, including mathematical reasoning, binary feedback, preference data, and demonstrations.

Beyond theory, we delve into infrastructure and practical implementation, showcasing how to efficiently evaluate IRL-based LLM alignment ideas in minutes. We conclude with insights from sparse-reward RL, covering reward shaping, credit assignment, and lessons from self-play.

By the end of this tutorial, attendees will gain a practical and theoretical understanding of LLM alignment through inverse RL, equipping them with the tools to build better-aligned models efficiently.

Turorial 2

T2: Eyetracking and NLP

David Reich, Omer Shubi, Lena Jäger and Yevgeni Berzak

This is a tutorial on the growing research area that combines eyetracking during reading with NLP. The tutorial will outline how eye movements in reading can be leveraged for NLP, and, vice versa, how NLP methods can advance psycholinguistic modeling of eye movements in reading. We will cover four main themes: (i) fundamentals of eye movements in reading, (ii) experimental methodologies and available data, (iii) integrating eye movement data in NLP models, and (iv) using NLP for modeling eye movements in reading. The tutorial is tailored to NLP researchers and practitioners, and will provide attendees with the essential background for conducting research on the joint modeling of eye movements and text.

Turorial 3

T3: Uncertainty Quantification for Large Language Models

Artem Shelmanov, Maxim Panov, Ekaterina Sergeevna Fadeeva, Artem Vazhentsev, Roman Konstantinovich Vashurin and Timothy Baldwin

In this tutorial, we cover foundational concepts of uncertainty quantification (UQ) for LLMs, present cutting-edge techniques, demonstrate practical applications of UQ in various tasks, and equip researchers and practitioners with tools for developing new UQ methods and harnessing uncertainty in various contexts.

Turorial 4

T4: Human-AI Collaboration: How AIs Augment Human Teammates

Tongshuang Wu, Diyi Yang, Kyle Lo and Marti A. Hearst

The continuous, rapid development of general- purpose models like LLMs suggests the theoretical possibility of AI performing any human task. Yet, despite the potential and promise, these models are far from perfect, excelling at certain tasks while struggling with others. The tension between what is possible and a model's limitations raises the general research question that has attracted attention from various disciplines: What is the best way to use AI to maximize its benefits? In this tutorial, we will review recent developments related to human-AI teaming and collaboration. To the best of our knowledge, our tutorial will be the first to provide a more integrated view from NLP, HCI, Computational Social Science, and Learning Science, etc., and highlight how different communities have identified the goals and societal impacts of such collaborations, both positive and negative. We will further discuss how to operationalize these Human-AI collaboration goals, and reflect on how state-of-the-art AI models should be evaluated and scaffolded to make them most useful in collaborative contexts.

Turorial 5

T5: Navigating Ethical Challenges in NLP: Hands-on strategies for students and researchers

Luciana Benotti, Fanny Ducel, Karen Fort, Guido Ivetta, Zhijing Jin, Min-Yen Kan, Seunghun Lee, Margot Mieskes, Minzhi Li, and Adriana Pagano

With NLP research being rapidly productionized into real-world applications, it is important to be aware of and think through the consequences of our research. Such ethical considerations are important in both authoring and reviewing (e.g. privacy, consent, fairness, etc).This tutorial will equip participants with ba-sic guidelines for thinking deeply about ethical issues and review common considerations that recur in NLP research. The methodology is interactive and participatory, including case studies and working in groups. Participants will gain practical experience on when to flag a paper for ethics review and how to write an ethical consideration section, which will be shared with the broader community. Importantly, the participants will be co-creating the tutorial outcomes and extending tutorial materials to share as public outcomes.

Turorial 6

T6: NLP for Counterspeech against Hate and Misinformation

Daniel Russo, Helena Bonaldi, Yi-Ling Chung,
Gavin Abercrombie and Marco Guerini

a. Introduction
b. Counterspeech in the real world: evidence from social science and the real world
c. NLP approaches to counterspeech against hate
d. NLP approaches to counterspeech for misinformation
e. Open challenges
f. Ethical considerations
g. Two guest talks from civil society/NGOs:
i. Cathy Buerger, Dangerous Speech Project ii. Simone Fontana, Facta News h. QA panel & Conclusion

Turorial 7

T7: Synthetic Data in the Era of Large Language Models

Vijay Viswanathan, Xiang Yue, Alisa Liu, Yizhong Wang and Graham Neubig

This tutorial unifies the rapid progress in synthetic data generation, with a focus on language data. We will discuss fundamental algorithms and major applications for training large language models when manually-curated data is limited.

Turorial 8

T8: Guardrails and Security for LLMs: Safe, Secure, and Controllable Steering of LLM Applications

Traian Rebedea, Leon Derczynski, Shaona Ghosh, Makesh Narsimhan Sreedhar, Faeze Brahman, Liwei Jiang, Bo Li, Yulia Tsvetkov, Christopher Parisien and Yejin Choi

Pretrained generative models, especially large language models, provide novel ways for users to interact with computers. While generative NLP research and applications had previously aimed at very domain-specific or task-specific solutions, current LLMs and applications (e.g. dialogue systems, agents) are versatile across many tasks and domains. Despite being trained to be helpful and aligned with human preferences (e.g., harmlessness), enforcing robust guardrails on LLMs remains a challenge. And, even when protected against rudimentary attacks, just like other complex software, LLMs can be vulnerable to attacks using sophisticated adversarial inputs. This tutorial provides a comprehensive overview of key guardrail mechanisms developed for LLMs, along with evaluation methodologies and a detailed security assessment protocol - including auto red-teaming of LLM-powered applications. Our aim is to move beyond the discussion of single prompt attacks and evaluation frameworks towards addressing how guardrailing can be done in complex dialogue systems that employ LLMs.

We aim to provide a cutting-edge and complete overview of deployment risks associated with LLMs in production environments. While the main focus will be on how to effectively protect against safety and security threats, we also tackle the more recent topic of providing dialogue and topical rails, including respecting custom policies. We also examine the new attack vectors introduced by LLM-enabled dialogue systems, such as methods for circumventing dialogue steering.

8

Main Conference

Main Conference: Monday, July 28, 2025

08:00–18:30	Registration	*Level 0 Entrance Hall*
09:00–09:30	**Session 1: Opening Session**	*Level 2 Hall A–C*
09:30–10:30	**Session 1: Keynote Luke Zettlemoyer**	*Level 2 Hall A–C*
10:30–11:00	*Coffee Break*	*Foyers & Hall X5*
11:00–12:30	***Session 2: Virtual Presentations***	*Underline*
11:00–12:30	**Session 2: In Person Posters (Main)**	
	NLP Applications	*Level 0 Exhibit Halls X4 or X5*
	Resources and Evaluation	*Level 0 Exhibit Halls X4 or X5*
	Multimodality and Language Grounding	*Level 0 Exhibit Halls X4 or X5*
	Efficient Methods for NLP	*Level 0 Exhibit Halls X4 or X5*
	Generation	*Level 0 Exhibit Halls X4 or X5*
	Ethics, Bias and Fairness	*Level 0 Exhibit Halls X4 or X5*
	Information Retrieval and Text Mining	*Level 0 Exhibit Halls X4 or X5*
	Human-centered NLP	*Level 0 Exhibit Halls X4 or X5*
	Speech Processing	*Level 0 Exhibit Halls X4 or X5*
12:30–14:00	*Lunch Break*	
14:00–15:30	**Session 3: In Person Orals**	
	Machine Translation	*Level 1 1.85*
	Semantics	*Level 1 1.86*
	Ethics, Bias, and Fairness 1	*Level 1 M.2*
	Generation	*Level 1 M.1*
	Human-centered NLP	*Level 1 1.62*
	Information Retrieval and Text Mining	*Level 1 N.1*
	Linguistic Theories and Cognitive Modeling	*Level 1 1.61*
	Multimodality and Language Grounding 1	*Level 2 Hall C*
	NLP Applications 1	*Level 2 Hall B*
	Resources and Evaluation 1	*Level 2 Hall A*

	Theme track: Generalisation of NLP models 1	*Level 1 1.15-16*
	Student Research Workshop	*Level 1 Hall N.2*
	Industry	*Level 1 Hall L.1 - L.3*
	BoF	*Level 1 1.31 - 1.32, 1.33*
15:30–16:00	*Coffee Break*	*Foyers & Hall X5*
16:00–16:30	***Session 4: CL Dissertation Award***	*Level 2 Hall A–C*
16:30–17:30	**Session 4: Panel**	
	Topic: Generalization on NLP Models	*Level 2 Hall A–C*
18:00–19:30	***Session 5: Virtual Presentations***	*Underline*
18:00–19:30	***Session 5: In Person Posters (Findings with Reception)***	*Level 0 Exhibit Halls X4 and X5*

Main Conference: Tuesday, July 29, 2025

08:30–21:30	Registration	*Level 0 Entrance Hall*
09:00–10:00	**Session 6: Keynote Verena Rieser**	*Level 2 Hall A–C*
10:00–10:30	*Coffee Break*	*Foyers & Hall X5*
10:30–12:00	***Session 7: Virtual Presentations***	*Underline*
10:30–12:00	**Session 7: In Person Posters (Main)**	
	NLP Applications	*Level 0 Exhibit Halls X4 or X5*
	Resources and Evaluation	*Level 0 Exhibit Halls X4 or X5*
	Language Modeling	*Level 0 Exhibit Halls X4 or X5*
	Interpretability and Model Analysis	*Level 0 Exhibit Halls X4 or X5*
	Question Answering	*Level 0 Exhibit Halls X4 or X5*
	Machine Learning for NLP	*Level 0 Exhibit Halls X4 or X5*
	Multilingualism and Cross-Lingual NLP	*Level 0 Exhibit Halls X4 or X5*
	Semantics	*Level 0 Exhibit Halls X4 or X5*
	Student Research Workshop	*Level 1 Hall N.2*
	Industry	*Level 1 Hall L.1–L.3*
	BoF	*Level 1 1.31–1.32, 1.33, 1.14*
12:00–13:00	*Lunch Break*	
13:00–13:45	***Session 8: Business Meeting***	*Level 2 Hall B*
13:45–14:00	***Break to Reset for orals***	
14:00–15:30	**Session 9: In Person Orals**	
	Dialogue and Interactive Systems	*Level 1 N.1*
	Interpretability and Model Analysis 1	*Level 1 M.2*
	Language Modeling 1	*Level 2 Hall C*
	Summarization and Information Extraction	*Level 1 N.2*
	Efficient Methods for NLP 1	*Level 1 M.1*
	Resources and Evaluation 2	*Level 2 Hall B*
	Phonology, Morphology, and Syntax	*Level 1 1.86*
	Sentiment, Style and Argumentation Analysis	*Level 1 1.62*
	Multilingualism and Cross-Lingual NLP	*Level 1 1.15-16*
	Speech Processing	*Level 1 1.61*
	Computational Social Science and Cultural Analytics	*Level 1 1.85*
	Industry	*Level 1 Hall L.1–L.3*
	BoF	*Level 1 1.31–1.32, 1.33, 1.14*
15:30–16:00	*Coffee Break*	*Foyers & Hall X5*
16:00–17:30	***Session 10: Virtual Presentations***	*Underline*
16:00–17:30	**Session 10: In Person Posters (Main)**	
	NLP Applications	*Level 0 Exhibit Halls X4 or X5*
	Resources and Evaluation	*Level 0 Exhibit Halls X4 or X5*
	Multimodality and Language Grounding	*Level 0 Exhibit Halls X4 or X5*
	Efficient Methods for NLP	*Level 0 Exhibit Halls X4 or X5*
	Generation	*Level 0 Exhibit Halls X4 or X5*
	Ethics, Bias, and Fairness	*Level 0 Exhibit Halls X4 or X5*
	Dialogue and Interactive Systems	*Level 0 Exhibit Halls X4 or X5*
	Theme track: Generalisation of NLP models	*Level 0 Exhibit Halls X4 or X5*

	Machine Translation	*Level 0 Exhibit Halls X4 or X5*
	Summarization and Information Extraction	*Level 0 Exhibit Halls X4 or X5*
	BoF	*Level 1 1.31–1.32, 1.33, 1.14*
19:00–22:00	***Social Event for all Registered Full, Main, or Social Event add-ons***	*Level 2 Hall A*

Main Conference: Wednesday, July 30, 2025

08:30–17:30	Registration	*Level 0 Entrance Hall*
09:00–10:30	**Session 11: In Person Orals**	
	Panel 1: Generalisation of NLP models	*Level 2 Hall A*
	Panel 2: LLM alignment	*Level 2 Hall B*
	Panel 3: Human-centred NLP	*Level 2 Hall C*
	NLP Applications 2	*Level 1 M.1*
	Multimodality and Language Grounding 2	*Level 1 M.2*
	Panel 4: Interpretability and model analysis	*Level 1 N.1*
	Panel 5: Multilinguality and language diversity	*Level 1 N.2*
	Machine Learning for NLP	*Level 1 1.61*
	Efficient Methods for NLP 3	*Level 1 1.62*
	Interpretability and Model Analysis 2	*Level 1 1.85*
	Question Answering	*Level 1 1.86*
	Language Modeling 2	*Level 1 1.15–1.16*
	Industry	*Level 1 Hall L.1–L.3*
	BoF	*Level 1 1.31–1.32, 1.33, 1.14*
10:30–11:00	*Coffee Break*	*Foyers & Hall X5*
11:00–12:30	***Session 12: Virtual Presentations***	*Underline*
11:00–12:30	**Session 12: In Person Posters (Main, IND, SRW)**	
	NLP Applications	*Level 0 Exhibit Halls X4 or X5*
	Resources and Evaluation	*Level 0 Exhibit Halls X4 or X5*
	Language Modeling	*Level 0 Exhibit Halls X4 or X5*
	Interpretability and Model Analysis	*Level 0 Exhibit Halls X4 or X5*
	Question Answering	*Level 0 Exhibit Halls X4 or X5*
	Computational Social Science and Cultural Analytics	*Level 0 Exhibit Halls X4 or X5*
	Linguistic Theories and Cognitive Modeling	*Level 0 Exhibit Halls X4 or X5*
	Sentiment, Style and Argumentation Analysis	*Level 0 Exhibit Halls X4 or X5*
	Phonology, Morphology, and Syntax	*Level 0 Exhibit Halls X4 or X5*
	Discourse and Pragmatics	*Level 0 Exhibit Halls X4 or X5*
	Student Research Workshop: Closing and Best Papers	*Level 1 Hall N.2*
	Industry: Car–Panel / Awards Closing	*Level 1 Hall L.1–L.3*
	BoF	*Level 1 1.31–1.32, 1.33, 1.14*
12:30–14:00	*Lunch Break*	
14:00–14:30	***Session 13: Presidential Address***	*Level 2 Hall A–C*
14:30–14:50	***Session 13: Test of Times***	*Level 2 Hall A–C*
14:50–15:00	***Session 13: ACL Distinguished Service Awards***	*Level 2 Hall A–C*
15:00–16:00	***Session 13: ACL Lifetime Achievement Award***	*Level 2 Hall A–C*
16:00–16:30	*Coffee Break*	*Hall 3 Lobby Level*
16:30–17:30	***Session 14: Best Paper Awards***	*Level 2 Hall A–C*
17:30–18:00	***Session 15: Closing Session***	*Level 2 Hall A–C*

9

Day 1: Monday 28 July

Oral Presentation

Session 2: Industry (Monday 28 July, 11:00-12:30, Hall L)

11:50 - 12:00

LLM-Enhanced Self-Evolving Reinforcement Learning for Multi-Step E-Commerce Payment Fraud Risk Detection

Bo Qu, Zhurong Wang, Daisuke Yagi, Zhen Xu, Yang Zhao, Yinan Shan, Frank Zahradnik

This paper presents a novel approach to e-commerce payment fraud detection by integrating reinforcement learning (RL) with Large Language Models (LLMs). By framing transaction risk as a multi-step Markov Decision Process (MDP), RL optimizes risk detection across multiple payment stages. Crafting effective reward functions, essential for RL model success, typically requires significant human expertise due to the complexity and variability in design. LLMs, with their advanced reasoning and coding capabilities, are well-suited to refine these functions, offering improvements over traditional methods. Our approach leverages LLMs to iteratively enhance reward functions, achieving better fraud detection accuracy and demonstrating zero-shot capability. Experiments with real-world data confirm the effectiveness, robustness, and resilience of our LLM-enhanced RL framework through long-term evaluations, underscoring the potential of LLMs in advancing industrial RL applications.

12:00 - 12:10

TN-Eval: Rubric and Evaluation Protocols for Measuring the Quality of Behavioral Therapy Notes

Raj Sanjay Shah, Lei Xu, Qianchu Liu, Jon Burnsky, Andrew Bertagnolli, Chaitanya Shivade

Behavioral therapy notes are important for both legal compliance and patient care. Unlike progress notes in physical health, quality standards for behavioral therapy notes remain underdeveloped. To address this gap, we collaborated with licensed therapists to design a comprehensive rubric for evaluating therapy notes across key dimensions: completeness, conciseness, and faithfulness. Further, we extend a public dataset of behavioral health conversations with therapist-written notes and LLM-generated notes, and apply our evaluation framework to measure their quality. We find that: (1) A rubric-based manual evaluation protocol offers more reliable and interpretable results than traditional Likert-scale annotations. (2) LLMs can mimic human evaluators in assessing completeness and conciseness but struggle with faithfulness. (3) Therapist-written notes often lack completeness and conciseness, while LLM-generated notes contain hallucinations. Surprisingly, in a blind test, therapists prefer and judge LLM-generated notes to be superior to therapist-written notes. As recruiting therapists for annotation is expensive, we will release the rubric, therapist-written notes, and expert annotations to support future research.

12:10 - 12:20

NeKo: Cross-Modality Post-Recognition Error Correction with Tasks-Guided Mixture-of-Experts Language Model

Yen-Ting Lin, Zhehuai Chen, Piotr Zelasko, Zhen Wan, Xuesong Yang, Zih-Ching Chen, Krishna C Puvvada, Ke Hu, Szu-Wei Fu, Jun Wei Chiu, Jagadeesh Balam, Boris Ginsburg, Yu-Chiang Frank Wang, Chao-Han Huck Yang

Construction of a general-purpose post-recognition error corrector poses a crucial question: how can we most effectively train a model on a large mixture of domain datasets? The answer would lie in learning dataset-specific features and digesting their knowledge in a single model. Previous methods achieve this by having separate correction language models, resulting in a significant increase in parameters. In this work, we present Mixture-of-Experts as a solution, highlighting that MoEs are much more than a scalability tool. We propose a Multi-Task Correction MoE, where we train the experts to become an ``expert'' of speech-to-text, language-to-text and vision-to-text datasets by learning to route each dataset's tokens to its mapped expert. Experiments on the Open ASR Leaderboard show that we explore a new state-of-the-art performance by achieving an average relative 5.0% WER reduction and substantial improvements in BLEU scores for speech and translation tasks. On zero-shot evaluation, NeKo outperforms GPT-3.5 and Claude-3.5-Sonnet with 15.5% to 27.6% relative WER reduction in the Hyporadise benchmark. NeKo performs competitively on grammar and post-OCR correction as a multi-task model.

12:20 - 12:30

Think Again! The Effect of Test-Time Compute on Preferences, Opinions, and Beliefs of Large Language Models

George Kour, Itay Nakash, Michal Shmueli-Scheuer, Ateret Anaby Tavor

As Large Language Models (LLMs) become deeply integrated into human life and increasingly influence decision-making, it's crucial to evaluate whether and to what extent they exhibit subjective preferences, opinions, and beliefs. These tendencies may stem from biases within the models, which may shape their behavior, influence the advice and recommendations they offer to users, and potentially reinforce certain viewpoints. This paper presents the Preference, Opinion, and Belief survey (POBs), a benchmark developed to assess LLMs' subjective inclinations across societal, cultural, ethical, and personal domains. We applied our benchmark to evaluate leading open- and closed-source LLMs, measuring desired properties such as reliability, neutrality, and consistency. In addition, we investigated the effect of increasing the test-time compute, through reasoning and self-reflection mechanisms, on those metrics. While effective in other tasks, our results show that these mechanisms offer only limited gains in our domain. Furthermore, we reveal that newer model versions are becoming less consistent and more biased toward specific viewpoints, highlighting a blind spot and a concerning trend. POBS: https://ibm.github.io/POBS

Session 3: CL Journal (Monday 28 July, 14:00-15:30, Hall M.2)

15:06 - 15:17

Large Language Models Are Biased Because They Are Large Language Models

Philip Resnik

This position paper's primary goal is to provoke thoughtful discussion about the relationship between bias and fundamental properties of large language models. I do this by seeking to convince the reader that harmful biases are an inevitable consequence arising from the design of any large language model as LLMs are currently formulated. To the extent that this is true, it suggests that the problem of harmful bias cannot be properly addressed without a serious reconsideration of AI driven by LLMs, going back to the foundational assumptions underlying their design.

Session 3: CL Journal (Monday 28 July, 14:00-15:30, Room 1.61)

15:17 - 15:28

Natural Language Processing RELIES on Linguistics

Juri Opitz Nathan Schneider Shira Wein

Large Language Models (LLMs) have become capable of generating highly fluent text in certain languages, without modules specially designed to capture grammar or semantic coherence. What does this mean for the future of linguistic expertise in NLP? We highlight several aspects in which NLP (still) relies on linguistics, or where linguistic thinking can illuminate new directions. We argue our case around the acronym RELIES, which encapsulates six major facets where linguistics contributes to NLP: Resources, Evaluation, Low-resource settings, Interpretability, Explanation, and the Study of language. This list is not exhaustive, nor is linguistics the main point of reference for every effort under these themes; but at a macro level, these facets highlight the enduring importance of studying machine systems vis-a-vis systems of human language.

Session 3: Ethics, Bias, and Fairness (Monday 28 July, 14:00-15:30, Hall M.2)

14:00 - 14:11

White Men Lead, Black Women Help? Benchmarking and Mitigating Language Agency Social Biases in LLMs

Yixin Wan, Kai-Wei Chang

Yixin Wan, Kai-Wei Chang

14:11 - 14:22

GenderAlign: An Alignment Dataset for Mitigating Gender Bias in Large Language Models

Tao Zhang, Ziqian Zeng, YuxiangXiao, Huiping Zhuang, Cen Chen, James R. Foulds, Shimei Pan

Tao Zhang, Ziqian Zeng, YuxiangXiao, Huiping Zhuang, Cen Chen, James R. Foulds, Shimei Pan

14:22 - 14:33

The Tug of War Within: Mitigating the Fairness-Privacy Conflicts in Large Language Models

Chen Qian, Dongrui Liu, Jie Zhang, Yong Liu, Jing Shao

Chen Qian, Dongrui Liu, Jie Zhang, Yong Liu, Jing Shao

14:33 - 14:44

Understanding Large Language Model Vulnerabilities to Social Bias Attacks

Jiaxu Zhao, Meng Fang, Fanghua Ye, Ke Xu, Qin Zhang, Joey Tianyi Zhou, Mykola Pechenizkiy

Jiaxu Zhao, Meng Fang, Fanghua Ye, Ke Xu, Qin Zhang, Joey Tianyi Zhou, Mykola Pechenizkiy

14:44 - 14:55

Amplifying Trans and Nonbinary Voices: A Community-Centred Harm Taxonomy for LLMs

Eddie L. Ungless, Sunipa Dev, Cynthia L. Bennett, Rebecca Gulotta, Jasmijn Bastings, Remi Denton

Eddie L. Ungless, Sunipa Dev, Cynthia L. Bennett, Rebecca Gulotta, Jasmijn Bastings, Remi Denton

14:55 - 15:06

Veracity Bias and Beyond: Uncovering LLMs' Hidden Beliefs in Problem-Solving Reasoning

Yue Zhou, Barbara Di Eugenio

Yue Zhou, Barbara Di Eugenio

Session 3: Generation (Monday 28 July, 14:00-15:30, Hall M.1)

14:00 - 14:11

Rethinking Repetition Problems of LLMs in Code Generation

Yihong Dong, Yuchen Liu, Xue Jiang, Zhi Jin, Ge Li

Yihong Dong, Yuchen Liu, Xue Jiang, Zhi Jin, Ge Li

14:11 - 14:22

Hierarchical Document Refinement for Long-context Retrieval-augmented Generation

Jiajie Jin, Xiaoxi Li, Guanting Dong, Yuyao Zhang, Yutao Zhu, Yongkang Wu, Zhonghua Li, YE QI, Zhicheng Dou

Jiajie Jin, Xiaoxi Li, Guanting Dong, Yuyao Zhang, Yutao Zhu, Yongkang Wu, Zhonghua Li, YE QI, Zhicheng Dou

14:22 - 14:33

APB: Accelerating Distributed Long-Context Inference by Passing Compressed Context Blocks across GPUs

Yuxiang Huang, Mingye Li, Xu Han, Chaojun Xiao, Weilin Zhao, Sun Ao, Hao Zhou, Jie Zhou, Zhiyuan Liu, Maosong Sun

Yuxiang Huang, Mingye Li, Xu Han, Chaojun Xiao, Weilin Zhao, Sun Ao, Hao Zhou, Jie Zhou, Zhiyuan Liu, Maosong Sun

14:33 - 14:44

Pre^3: Enabling Deterministic Pushdown Automata for Faster Structured LLM Generation

Junyi Chen, Shihao Bai, Zaijun Wang, Siyu Wu, Chuheng Du, Hailong Yang, Ruihao Gong, Shengzhong Liu, Fan Wu, Guihai Chen

Junyi Chen, Shihao Bai, Zaijun Wang, Siyu Wu, Chuheng Du, Hailong Yang, Ruihao Gong, Shengzhong Liu, Fan Wu, Guihai Chen

14:44 - 14:55

TC–RAG: Turing–Complete RAG's Case study on Medical LLM Systems

Xinke Jiang, Yue Fang, Rihong Qiu, Haoyu Zhang, Yongxin Xu, Hao Chen, WentaoZhang, Ruizhe Zhang, Yuchen Fang, Xinyu Ma, Xu Chu, Junfeng Zhao, Yasha Wang

Xinke Jiang, Yue Fang, Rihong Qiu, Haoyu Zhang, Yongxin Xu, Hao Chen, WentaoZhang, Ruizhe Zhang, Yuchen Fang, Xinyu Ma, Xu Chu, Junfeng Zhao, Yasha Wang

14:55 - 15:06

PlanGenLLMs: A Modern Survey of LLM Planning Capabilities

Hui Wei, Zihao Zhang, Shenghua He, Tian Xia, Shijia Pan, Fei Liu

Hui Wei, Zihao Zhang, Shenghua He, Tian Xia, Shijia Pan, Fei Liu

15:06 - 15:17

T2A-Feedback: Improving Basic Capabilities of Text-to-Audio Generation via Fine-grained AI Feedback

Zehan Wang, Ke Lei, Chen Zhu, Jiawei Huang, Sashuai zhou, Luping Liu, Xize Cheng, Shengpeng Ji, Zhenhui Ye, Tao Jin, Zhou Zhao

Zehan Wang, Ke Lei, Chen Zhu, Jiawei Huang, Sashuai zhou, Luping Liu, Xize Cheng, Shengpeng Ji, Zhenhui Ye, Tao Jin, Zhou Zhao

15:17 - 15:28

LLMs know their vulnerabilities: Uncover Safety Gaps through Natural Distribution Shifts

Qibing Ren, Hao Li, Dongrui Liu, Zhanxu Xie, Xiaoya Lu, Yu Qiao, Lei Sha, Junchi Yan, Lizhuang Ma, Jing Shao

Qibing Ren, Hao Li, Dongrui Liu, Zhanxu Xie, Xiaoya Lu, Yu Qiao, Lei Sha, Junchi Yan, Lizhuang Ma, Jing Shao

Session 3: Human-Centered NLP (Monday 28 July, 14:00-15:30, Room 1.62)

14:00 - 14:11

Algorithmic Fidelity of Large Language Models in Generating Synthetic German Public Opinions: A Case Study

Bolei Ma, Berk Yoztyurk, Anna-Carolina Haensch, Xinpeng Wang, Markus Herklotz, Frauke Kreuter, Barbara Plank, Matthias Aßenmacher

Bolei Ma, Berk Yoztyurk, Anna-Carolina Haensch, Xinpeng Wang, Markus Herklotz, Frauke Kreuter, Barbara Plank, Matthias Aßenmacher

14:11 - 14:22

Biased LLMs can Influence Political Decision-Making

Jillian Fisher, Shangbin Feng, Robert Aron, Thomas Richardson, Yejin Choi, Daniel W Fisher, Jennifer Pan, Yulia Tsvetkov, Katharina Reinecke

Jillian Fisher, Shangbin Feng, Robert Aron, Thomas Richardson, Yejin Choi, Daniel W Fisher, Jennifer Pan, Yulia Tsvetkov, Katharina Reinecke

14:22 - 14:33

GRACE: A Granular Benchmark for Evaluating Model Calibration against Human Calibration

Yoo Yeon Sung, Eve Fleisig, Yu Hou, Ishan Upadhyay, Jordan Lee Boyd-Graber

Yoo Yeon Sung, Eve Fleisig, Yu Hou, Ishan Upadhyay, Jordan Lee Boyd-Graber

14:33 - 14:44

Dehumanizing Machines: Mitigating Anthropomorphic Behaviors in Text Generation Systems

Myra Cheng, Su Lin Blodgett, Alicia DeVrio, Lisa Egede, Alexandra Olteanu

Myra Cheng, Su Lin Blodgett, Alicia DeVrio, Lisa Egede, Alexandra Olteanu

14:44 - 14:55

Evaluating Multimodal Language Models as Visual Assistants for Visually Impaired Users

Antonia Karamolegkou, Malvina Nikandrou, Georgios Pantazopoulos, Danae Sanchez Villegas, Phillip Rust, Ruchira Dhar, Daniel Hershcovich, Anders Søgaard

Antonia Karamolegkou, Malvina Nikandrou, Georgios Pantazopoulos, Danae Sanchez Villegas, Phillip Rust, Ruchira Dhar, Daniel Hershcovich, Anders Søgaard

14:55 - 15:06

The Esethu Framework: Reimagining Sustainable Dataset Governance and Curation for Low-Resource Languages

Jenalea Rajab, Anuoluwapo Aremu, Everlyn Asiko Chimoto, Graham Morrissey, Fadel Thior, Jessica Ojo, Atnafu Lambebo Tonja, Wilhelmina Nekoto, Pelonomi Moiloa, Jade Abbott, Vukosi Marivate, Benjamin Rosman

Jenalea Rajab, Anuoluwapo Aremu, Everlyn Asiko Chimoto, Graham Morrissey, Fadel Thior, Jessica Ojo, Atnafu Lambebo Tonja, Wilhelmina Nekoto, Pelonomi Moiloa, Jade Abbott, Vukosi Marivate, Benjamin Rosman

15:06 - 15:17

Persona Dynamics: Unveiling the Impact of Persona Traits on Agents in Text-Based Games

Seungwon Lim, Seungbeen Lee, Dongjun Min, Youngjae Yu

Seungwon Lim, Seungbeen Lee, Dongjun Min, Youngjae Yu

15:17 - 15:28

Comparison-based Active Preference Learning for Multi-dimensional Personalization

Minhyeon Oh, Seungjoon Lee, Jungseul Ok

Minhyeon Oh, Seungjoon Lee, Jungseul Ok

Session 3: Industry (Monday 28 July, 14:00-15:30, Hall L)

14:00 - 14:10

MathAgent: Leveraging a Mixture-of-Math-Agent Framework for Real-World Multimodal Mathematical Error Detection

Yibo Yan, Shen Wang, Jiahao Huo, Philip S. Yu, Xuming Hu, Qingsong Wen

Mathematical error detection in educational settings presents a significant challenge for Multimodal Large Language Models (MLLMs), requiring a sophisticated understanding of both visual and textual mathematical content along with complex reasoning capabilities. Though effective in mathematical problem-solving, MLLMs often struggle with the nuanced task of **identifying and categorizing student errors in multimodal mathematical contexts**. Therefore, we introduce **MathAgent, a novel Mixture-of-Math-Agent framework** specifically designed to address these challenges. Our approach decomposes error detection into three phases with specialized agents: an image-text consistency validator, a visual semantic interpreter, and an integrative error analyzer. This architecture enables more accurate processing of multimodal mathematical content by explicitly modeling the relationships between multimodal problems and student solution steps. We evaluate MathAgent on real-world educational data, demonstrating approximately 5% higher accuracy in error step identification and 3% improvement in error categorization compared to baseline models. Furthermore, MathAgent has been successfully deployed in an educational platform serving over one million K-12 students, achieving nearly 90% student satisfaction while generating significant cost savings by reducing manual error detection.

14:10 - 14:20

Run LoRA Run: Faster and Lighter LoRA Implementations

Daria Cherniuk, Aleksandr Mikhalev, Ivan Oseledets

LoRA is a technique that reduces the number of trainable parameters in a neural network by introducing low-rank adapters to linear layers. This technique is used for fine-tuning and even training large transformer models from scratch. This paper presents the RunLoRA framework for efficient implementations of LoRA, which significantly improves the speed of neural network training and fine-tuning with low-rank adapters. The proposed implementation optimizes the computation of LoRA operations based on the shape of the corresponding linear layer weights, the input dimensions, and the LoRA rank by selecting the best forward and backward computation graphs based on FLOPs and time estimations. This results in faster training without sacrificing accuracy. The experimental results show a speedup ranging from 10% to 28% on various transformer models.

14:20 - 14:30

HierGR: Hierarchical Semantic Representation Enhancement for Generative Retrieval in Food Delivery Search

Fuwei Zhang, Xiaoyu Liu, Jia Xinyu, Yingfei Zhang, Zenghua Xia, Fei Jiang, Fuzhen Zhuang, Wei Lin, Zhao Zhang

Food delivery search aims to quickly retrieve deliverable items that meet users' needs, typically requiring faster and more accurate query understanding compared to traditional e-commerce search. Generative retrieval (GR), an emerging search paradigm, harnesses the advanced query understanding capabilities of large language models (LLMs) to enhance the retrieval of results for complex and long-tail queries in food delivery search scenarios. However, there are still challenges in deploying GR to online scenarios: 1) **the large scale of items**; 2) **latency constraints unmet by LLM inference in online retrieval**; and 3) **strong location-based service restrictions on generated items**. To explore the application of GR in food delivery search, we optimize both offline training and online deployment, proposing **Hier**archical semantic representation enhancement for **G**enerative **R**etrieval (HierGR). Specifically, for the generation of semantic IDs, we propose an optimization method that refines the residual quantization process to generate hierarchically semantic IDs for items. Additionally, to successfully deploy on a well-known food delivery platform, we utilize the query cache mechanism and integrate the GR model with the online dense retrieval model to fulfill real-world search requirements. Online A/B testing results show that our proposed method increases **the number of online orders by 0.68\%** for complex search intents. The source code is available at https://github.com/zhangfw123/HierGR.

14:30 - 14:40

Generating Q\&A Benchmarks for RAG Evaluation in Enterprise Settings

Simone Filice, Guy Horowitz, David Carmel, Zohar Karnin, Liane Lewin-Eytan, Yoelle Maarek

We introduce DataMorgana, a tool for generating synthetic Q\&A benchmarks tailored to RAG applications in enterprise settings. DataMorgana enables customization of the generated benchmark according to the expected diverse traffic of the RAG application. It allows for specifying question types and their associated distribution via a lightweight configuration mechanism. We demonstrate via a series of quantitative and qualitative experiments that DataMorgana surpasses existing tools in terms of lexical, syntactic, and semantic diversity of the generated benchmark while maintaining high quality. We run our experiments over domain-specific and general-knowledge public datasets, as well as two private datasets from governmental RAG applications: one for citizens and the other for government employees. The private datasets have been shared with us by AI71, an AI company, which has integrated DataMorgana into its offerings. In addition, DataMorgana has been offered to about 150 researchers worldwide as part of the SIGIR'2025 LiveRAG Challenge held in Spring 2025.

14:40 - 14:50

To Chat or Task: a Multi-turn Dialogue Generation Framework for Task-Oriented Dialogue Systems

Daniel Rim, Minsoo Cho, Changwoo Chun, Jaegul Choo

Task-oriented dialogue systems employ natural language understanding (NLU) modules to manage the intricate and continually evolving business requirements of production systems. Although the development of Large Language Models (LLMs) introduced extraordinary chitchat capabilities, implementing LLMs into such systems brought new difficulties. One of the main challenges is the lack of specific datasets for training and evaluation of systems that offer both capabilities: chat and task. As NLU modules are designed to handle complex task requests and LLMs are utilized to specifically answer chitchat interactions, the system must correctly identify the functional intent of the user to utilize an applicable module. This paper presents CTFusion, a multi-turn dialogue generation framework designed to assist the evaluation and training of production systems that offer both capabilities. Utilizing the framework, we generate a multi-turn dialogue dataset for in-vehicle speech recognition system, which includes 41,211 dialogues of 240 real-world in-vehicle intents, and train In-vehicle Context Sensor (ICS), a lightweight model that successfully identifies the functional intent of the driver. ICS outperforms all baseline models across various experimental settings, which demonstrates that CTFusion can help generate relevant datasets with a complex business logic, which can subsequently assist production systems in leveraging LLMs for their chitchat capabilities.

14:50 - 15:00

Scaling Under-Resourced TTS: A Data-Optimized Framework with Advanced Acoustic Modeling for Thai

Yizhong Geng, Jizhuo Xu, Zeyu Liang, Jinghan Yang, Xiaoyi Shi, Xiaoyu Shen

Text-to-speech (TTS) technology has achieved impressive results for widely spoken languages, yet many under-resourced languages remain challenged by limited data and linguistic complexities. In this paper, we present a novel methodology that integrates a data-optimized framework with an advanced acoustic model to build high-quality TTS systems for low-resource scenarios. We demonstrate the effectiveness of our approach using Thai as an illustrative case, where intricate phonetic rules and sparse resources are effectively addressed. Our method enables zero-shot voice cloning and improved performance across diverse client applications, ranging from finance to healthcare, education, and law. Extensive evaluations—both subjective and objective—confirm that our model meets state-of-the-art standards, offering a scalable solution for TTS production in data-limited settings, with significant implications for broader industry adoption and multilingual accessibility. All demos are available in .

15:00 - 15:10

Speed Without Sacrifice: Fine-Tuning Language Models with Medusa and Knowledge Distillation in Travel Applications

Daniel Zagyva, Emmanouil Stergiadis, Laurens van der Maas, Aleksandra Dokic, Eran Fainman, Ilya Gusev, Moran Beladev

In high-stakes industrial NLP applications, balancing generation quality with speed and efficiency presents significant challenges. We address them by investigating two complementary optimization approaches: Medusa for speculative decoding and knowledge distillation (KD) for model compression. We demonstrate the practical application of these techniques in real-world travel domain tasks, including trip planning, smart filters, and generating accommodation descriptions. We introduce modifications to the Medusa implementation, starting with base pre-trained models rather than conversational fine-tuned ones, and developing a simplified single-stage training process for Medusa-2 that maintains performance while reducing computational requirements. Lastly, we present a novel framework that combines Medusa with knowledge distillation, achieving compounded benefits in both model size and inference speed. Our experiments with TinyLlama-1.1B as the student model and Llama-3.1-70B as the teacher show that the combined approach maintains the teacher's performance quality while reducing inference latency by 10-20x.

15:10 - 15:20

Proactive Guidance of Multi-Turn Conversation in Industrial Search

Xiaoyu Li, Xiao Li, Li Gao, Yiding Liu, Xiaoyang Wang, Shuaiqiang Wang, Junfeng Wang, Dawei Yin

The evolution of Large Language Models (LLMs) has significantly advanced multi-turn conversation systems, emphasizing the need for proactive guidance to enhance users' interactions. However, these systems face challenges in dynamically adapting to shifts in users' goals and maintaining low latency for real-time interactions. In the Baidu Search AI assistant, an industrial-scale multi-turn search system, we propose a novel two-phase framework to provide proactive guidance. The first phase, Goal-adaptive Supervised Fine-Tuning (G-SFT), employs a goal adaptation agent that dynamically adapts to user goal shifts and provides goal-relevant contextual information. G-SFT also incorporates scalable knowledge transfer to distill insights from LLMs into a lightweight model for real-time interaction. The second phase, Click-oriented Reinforcement Learning (C-RL), adopts a generate-rank paradigm, systematically constructs preference pairs from user click signals, and proactively improves click-through rates through more engaging guidance. This dual-phase architecture achieves complementary objectives: G-SFT ensures accurate goal tracking, while C-RL optimizes interaction quality through click signal-driven reinforcement learning. Extensive experiments demonstrate that our framework achieves 86.10\% accuracy in offline evaluation (+23.95\% over baseline) and 25.28\% CTR in online deployment (149.06\% relative improvement), while reducing inference latency by 69.55\% through scalable knowledge distillation.

15:20 - 15:30

SpeechWeave: Diverse Multilingual Synthetic Text \& Audio Data Generation Pipeline for Training Text to Speech Models

Karan Dua, Puneet Mittal, Ranjeet Gupta, Hitesh Laxmichand Patel

High-quality Text-to-Speech (TTS) model training requires extensive and diverse text and speech data. It is challenging to procure such data from real sources due to issues of domain specificity, licensing, and scalability. Large language models (LLMs) can certainly generate textual data, but they create repetitive text with insufficient variation in the prompt during the generation process. Another important aspect in TTS training data is text normalization. Tools for normalization might occasionally introduce anomalies or overlook valuable patterns, and thus impact data quality. Furthermore, it is also impractical to rely on voice artists for large scale speech recording in commercial TTS systems with standardized voices. To address these challenges, we propose SpeechWeave, a synthetic speech data generation pipeline that is capable of automating the generation of multilingual, domain-specific datasets for training TTS models. Our experiments reveal that our pipeline generates data that is 10–48% more diverse than the baseline across various linguistic and phonetic metrics, along with speaker-standardized speech audio while generating approximately 97% correctly normalized text. Our approach enables scalable, high-quality data generation for TTS training, improving diversity, normalization, and voice consistency in the generated datasets.

Session 3: Information Retrieval and Text Mining (Monday 28 July, 14:00-15:30, Hall N.1)

14:00 - 14:11

Sliding Windows Are Not the End: Exploring Full Ranking with Long-Context Large Language Models

Wenhan Liu, Xinyu Ma, Yutao Zhu, Ziliang Zhao, Shuaiqiang Wang, Dawei Yin, Zhicheng Dou

Wenhan Liu, Xinyu Ma, Yutao Zhu, Ziliang Zhao, Shuaiqiang Wang, Dawei Yin, Zhicheng Dou

14:11 - 14:22

MAPS: Motivation-Aware Personalized Search via LLM-Driven Consultation Alignment
Weicong Qin, Yi Xu, Weijie Yu, Chenglei Shen, Ming He, Jianping Fan, Xiao Zhang, Jun Xu
Weicong Qin, Yi Xu, Weijie Yu, Chenglei Shen, Ming He, Jianping Fan, Xiao Zhang, Jun Xu

14:22 - 14:33

QAEncoder: Towards Aligned Representation Learning in Question Answering Systems
Zhengren Wang, Qinhan Yu, Shida Wei, Zhiyu li, Feiyu Xiong, Xiaoxing Wang, Simin Niu, Hao Liang, Wentao Zhang
Zhengren Wang, Qinhan Yu, Shida Wei, Zhiyu li, Feiyu Xiong, Xiaoxing Wang, Simin Niu, Hao Liang, Wentao Zhang

14:33 - 14:44

MegaPairs: Massive Data Synthesis for Universal Multimodal Retrieval
Junjie Zhou, yongping xiong, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, Defu Lian
Junjie Zhou, yongping xiong, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, Defu Lian

14:44 - 14:55

Reinforced IR: A Dual Reinforcement Framework For Domain-Adapted Information Retrieval
Chaofan Li, Jianlyu Chen, Yingxia Shao, Chaozhuo Li, Quanqing Xu, Defu Lian, Zheng Liu
Chaofan Li, Jianlyu Chen, Yingxia Shao, Chaozhuo Li, Quanqing Xu, Defu Lian, Zheng Liu

14:55 - 15:06

Don't Reinvent the Wheel: Efficient Instruction-Following Text Embedding based on Guided Space Transformation
Yingchaojie Feng, Yiqun Sun, Yandong Sun, Minfeng Zhu, Qiang Huang, Anthony Kum Hoe Tung, Wei Chen
Yingchaojie Feng, Yiqun Sun, Yandong Sun, Minfeng Zhu, Qiang Huang, Anthony Kum Hoe Tung, Wei Chen

15:06 - 15:17

PRISM: A Framework for Producing Interpretable Political Bias Embeddings with Political-Aware Cross-Encoder
Yiqun Sun, Qiang Huang, Anthony Kum Hoe Tung, Jun Yu
Yiqun Sun, Qiang Huang, Anthony Kum Hoe Tung, Jun Yu

15:17 - 15:28

IMOL: Incomplete-Modality-Tolerant Learning for Multi-Domain Fake News Video Detection
Zhi Zeng, Jiaying Wu, Minnan Luo, Herun Wan, Xiangzheng Kong, Zihan Ma, Guang Dai, Qinghua Zheng
Zhi Zeng, Jiaying Wu, Minnan Luo, Herun Wan, Xiangzheng Kong, Zihan Ma, Guang Dai, Qinghua Zheng

Session 3: Linguistic theories, Cognitive Modeling and Psycholinguistics (Monday 28 July, 14:00-15:30, Room 1.61)

14:00 - 14:11

The Impact of Token Granularity on the Predictive Power of Language Model Surprisal

Byung-Doh Oh, William Schuler

Byung-Doh Oh, William Schuler

14:11 - 14:22

Rolling the DICE on Idiomaticity: How LLMs Fail to Grasp Context

Maggie Mi, Aline Villavicencio, Nafise Sadat Moosavi

Maggie Mi, Aline Villavicencio, Nafise Sadat Moosavi

14:22 - 14:33

When the LM misunderstood the human chuckled: Analyzing garden path effects in humans and language models

Samuel Joseph Amouyal, Aya Meltzer-Asscher, Jonathan Berant

Samuel Joseph Amouyal, Aya Meltzer-Asscher, Jonathan Berant

14:33 - 14:44

Growing Through Experience: Scaling Episodic Grounding in Language Models

Chunhui Zhang, Sirui Wang, Zhongyu Ouyang, Xiangchi Yuan, Soroush Vosoughi

Chunhui Zhang, Sirui Wang, Zhongyu Ouyang, Xiangchi Yuan, Soroush Vosoughi

14:44 - 14:55

Logical forms complement probability in understanding language model (and human) performance

Yixuan Wang, Freda Shi

Yixuan Wang, Freda Shi

14:55 - 15:06

Déjà Vu? Decoding Repeated Reading from Eye Movements

Yoav Meiri, Omer Shubi, Cfir Avraham Hadar, Ariel Kreisberg Nitzav, Yevgeni Berzak

Yoav Meiri, Omer Shubi, Cfir Avraham Hadar, Ariel Kreisberg Nitzav, Yevgeni Berzak

15:06 - 15:17

Using Information Theory to Characterize Prosodic Typology: The Case of Tone, Pitch-Accent and Stress-Accent

Ethan Wilcox, Cui Ding, Giovanni Acampa, Tiago Pimentel, Alex Warstadt, Tamar I Regev

Ethan Wilcox, Cui Ding, Giovanni Acampa, Tiago Pimentel, Alex Warstadt, Tamar I Regev

Session 3: Machine Translation (Monday 28 July, 14:00-15:30, Room1.85)

14:00 - 14:11

CECT dataset: Overcoming Data Scarcity in Context-Aware E-Commerce MT

Mikołaj Pokrywka, Wojciech Kusa, Mieszko Rutkowski, Mikołaj Koszowski

Mikołaj Pokrywka, Wojciech Kusa, Mieszko Rutkowski, Mikołaj Koszowski

14:11 - 14:22

Unifying Continuous and Discrete Text Diffusion with Non-simultaneous Diffusion Processes

Bocheng Li, Zhujin Gao, Linli Xu

Bocheng Li, Zhujin Gao, Linli Xu

14:22 - 14:33

MLAS-LoRA: Language-Aware Parameters Detection and LoRA-Based Knowledge Transfer for Multilingual Machine Translation

Tianyu Dong, Bo Li, Jinsong Liu, shaolin Zhu, Deyi Xiong

Tianyu Dong, Bo Li, Jinsong Liu, shaolin Zhu, Deyi Xiong

14:33 - 14:44

Improving Language and Modality Transfer in Translation by Character-level Modeling

Ioannis Tsiamas, David Dale, Marta R. Costa-jussà

Ioannis Tsiamas, David Dale, Marta R. Costa-jussà

14:44 - 14:55

Registering Source Tokens to Target Language Spaces in Multilingual Neural Machine Translation

Zhi Qu, Yiran Wang, Jiannan Mao, Chenchen Ding, Hideki Tanaka, Masao Utiyama, Taro Watanabe

Zhi Qu, Yiran Wang, Jiannan Mao, Chenchen Ding, Hideki Tanaka, Masao Utiyama, Taro Watanabe

14:55 - 15:06

Has Machine Translation Evaluation Achieved Human Parity? The Human Reference and the Limits of Progress

Lorenzo Proietti, Stefano Perrella, Roberto Navigli

Lorenzo Proietti, Stefano Perrella, Roberto Navigli

15:06 - 15:17

Make Imagination Clearer! Stable Diffusion-based Visual Imagination for Multimodal Machine Translation

Andong Chen, Yuchen Song, Kehai Chen, Xuefeng Bai, Muyun Yang, Liqiang Nie, Jie Liu, Tiejun Zhao, Min zhang

Andong Chen, Yuchen Song, Kehai Chen, Xuefeng Bai, Muyun Yang, Liqiang Nie, Jie Liu, Tiejun Zhao, Min zhang

15:17 - 15:28

It's Not a Walk in the Park! Challenges of Idiom Translation in Speech-to-text Systems

Iuliia Zaitova, Badr M. Abdullah, Wei Xue, Dietrich Klakow, Bernd Möbius, Tania Avgustinova

Iuliia Zaitova, Badr M. Abdullah, Wei Xue, Dietrich Klakow, Bernd Möbius, Tania Avgustinova

Session 3: Multimodality and Language Grounding to Vision, Robotics and Beyond (Monday 28 July, 14:00-15:30, Hall C)

14:00 - 14:11

Can Multimodal Large Language Models Understand Spatial Relations?

Jingping Liu, Ziyan Liu, Zhedong Cen, Yan Zhou, Yinan Zou, Weiyan Zhang, Haiyun Jiang, Tong Ruan

Jingping Liu, Ziyan Liu, Zhedong Cen, Yan Zhou, Yinan Zou, Weiyan Zhang, Haiyun Jiang, Tong Ruan

14:11 - 14:22

In-the-wild Audio Spatialization with Flexible Text-guided Localization

Tianrui Pan, Jie Liu, Zewen Huang, Jie Tang, Gangshan Wu

Tianrui Pan, Jie Liu, Zewen Huang, Jie Tang, Gangshan Wu

14:22 - 14:33

mPLUG-DocOwl2: High-resolution Compressing for OCR-free Multi-page Document Understanding

Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, Jingren Zhou

Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, Jingren Zhou

14:33 - 14:44

OS Agents: A Survey on MLLM-based Agents for Computer, Phone and Browser Use

Xueyu Hu, Tao Xiong, Biao Yi, Ruixuan Xiao, Zishu Wei, Yurun Chen, Jiasheng Ye, Meiling Tao, Xiangxin Zhou, Ziyu Zhao, Yuhuai Li, Shengze Xu, Shenzhi Wang, Xinchen Xu, Shuofei Qiao, Zhaokai Wang, Kun Kuang, Tieyong Zeng, Liang Wang, Jiwei Li, Yuchen Eleanor Jiang, Wangchunshu Zhou, Guoyin Wang, Keting Yin, Zhou Zhao, Hongxia Yang, Fan Wu, Shengyu Zhang, Fei Wu

Xueyu Hu, Tao Xiong, Biao Yi, Ruixuan Xiao, Zishu Wei, Yurun Chen, Jiasheng Ye, Meiling Tao, Xiangxin Zhou, Ziyu Zhao, Yuhuai Li, Shengze Xu, Shenzhi Wang, Xinchen Xu, Shuofei Qiao, Zhaokai Wang, Kun Kuang, Tieyong Zeng, Liang Wang, Jiwei Li, Yuchen Eleanor Jiang, Wangchunshu Zhou, Guoyin Wang, Keting Yin, Zhou Zhao, Hongxia Yang, Fan Wu, Shengyu Zhang, Fei Wu

14:44 - 14:55

SpaRE: Enhancing Spatial Reasoning in Vision-Language Models with Synthetic Data

Michael Ogezi, Freda Shi

Michael Ogezi, Freda Shi

14:55 - 15:06

CADReview: Automatically Reviewing CAD Programs with Error Detection and Correction

Jiali Chen, Xusen Hei, HongFei Liu, Yuancheng Wei, Zikun Deng, Jiayuan Xie, Yi Cai, Li Qing

Jiali Chen, Xusen Hei, HongFei Liu, Yuancheng Wei, Zikun Deng, Jiayuan Xie, Yi Cai, Li Qing

15:06 - 15:17

The Role of Visual Modality in Multimodal Mathematical Reasoning: Challenges and Insights

Yufang Liu, Yao Du, Tao Ji, Jianing Wang, Yang Liu, Yuanbin Wu, Aimin Zhou, Mengdi Zhang, Xunliang Cai

Yufang Liu, Yao Du, Tao Ji, Jianing Wang, Yang Liu, Yuanbin Wu, Aimin Zhou, Mengdi Zhang, Xunliang Cai

15:17 - 15:28

I0T: Embedding Standardization Method Towards Zero Modality Gap

Na Min An, Eunki Kim, James Thorne, Hyunjung Shim

Na Min An, Eunki Kim, James Thorne, Hyunjung Shim

Session 3: NLP Applications (Monday 28 July, 14:00-15:30, Hall B)

14:00 - 14:11

Disentangling Biased Knowledge from Reasoning in Large Language Models via Machine Unlearning

Zheyuan Liu, Suraj Maharjan, Fanyou Wu, Rahil Parikh, Belhassen Bayar, Srinivasan H. Sengamedu, Meng Jiang

Zheyuan Liu, Suraj Maharjan, Fanyou Wu, Rahil Parikh, Belhassen Bayar, Srinivasan H. Sengamedu, Meng Jiang

14:11 - 14:22

HD-NDEs: Neural Differential Equations for Hallucination Detection in LLMs

Qing Li, Jiahui Geng, Zongxiong Chen, Derui Zhu, Yuxia Wang, Congbo Ma, Chenyang Lyu, Fakhri Karray

Qing Li, Jiahui Geng, Zongxiong Chen, Derui Zhu, Yuxia Wang, Congbo Ma, Chenyang Lyu, Fakhri Karray

14:22 - 14:33

Evaluation Agent: Efficient and Promptable Evaluation Framework for Visual Generative Models

Fan Zhang, Shulin Tian, Ziqi Huang, Yu Qiao, Ziwei Liu

Fan Zhang, Shulin Tian, Ziqi Huang, Yu Qiao, Ziwei Liu

14:33 - 14:44

Context Matters: Semantic Expansion and Taxonomy-Grounded Augmentation for Sexism Detection

SAHRISH KHAN, Gabriele Pergola, Arshad Jhumka

SAHRISH KHAN, Gabriele Pergola, Arshad Jhumka

14:44 - 14:55

Do not Abstain! Identify and Solve the Uncertainty

Jingyu Liu, JingquanPeng, xiaopeng Wu, Xubin Li, Tiezheng Ge, Bo Zheng, Yong Liu

Jingyu Liu, JingquanPeng, xiaopeng Wu, Xubin Li, Tiezheng Ge, Bo Zheng, Yong Liu

14:55 - 15:06

Generate First, Then Sample: Enhancing Fake News Detection with LLM-Augmented Reinforced Sampling

Zhao Tong, Yimeng Gu, Huidong Liu, Qiang Liu, Shu Wu, Haichao Shi, Xiao-Yu Zhang

Zhao Tong, Yimeng Gu, Huidong Liu, Qiang Liu, Shu Wu, Haichao Shi, Xiao-Yu Zhang

15:06 - 15:17

SDD: Self-Degraded Defense against Malicious Fine-tuning

ZiXuan Chen, Weikai Lu, Xin Lin, Ziqian Zeng

ZiXuan Chen, Weikai Lu, Xin Lin, Ziqian Zeng

15:17 - 15:28

Introducing Verification Task of Set Consistency with Set-Consistency Energy Networks

Mooho Song, Hye Ryung Son, Jay-Yoon Lee

Mooho Song, Hye Ryung Son, Jay-Yoon Lee

Session 3: Resources and Evaluation (Monday 28 July, 14:00-15:30, Hall A)

14:00 - 14:11

AfriMed-QA: A Pan-African, Multi-Specialty, Medical Question-Answering Benchmark Dataset

Charles Nimo, Tobi Olatunji, Abraham Toluwase Owodunni, Tassallah Abdullahi, Emmanuel Ayodele, Mardhiyah Sanni, Ezinwanne C. Aka, Folafunmi Omofoye, Foutse Yuehgoh, Timothy Faniran, Bonaventure F. P. Dossou, Moshood O. Yekini, Jonas Kemp, Katherine A Heller, Jude Chidubem Omeke, Chidi Asuzu MD, Naome A Etori, Aïmérou Ndiaye, Ifeoma Okoh, Evans Doe Ocansey, Wendy Kinara, Michael Best, Irfan Essa, Stephen Edward Moore, Chris Fourie, Mercy Nyamewaa Asiedu

Charles Nimo, Tobi Olatunji, Abraham Toluwase Owodunni, Tassallah Abdullahi, Emmanuel Ayodele, Mardhiyah Sanni, Ezinwanne C. Aka, Folafunmi Omofoye, Foutse Yuehgoh, Timothy Faniran, Bonaventure F. P. Dossou, Moshood O. Yekini, Jonas Kemp, Katherine A Heller, Jude Chidubem Omeke, Chidi Asuzu MD, Naome A Etori, Aïmérou Ndiaye, Ifeoma Okoh, Evans Doe Ocansey, Wendy Kinara, Michael Best, Irfan Essa, Stephen Edward Moore, Chris Fourie, Mercy Nyamewaa Asiedu

14:11 - 14:22

HateDay: Insights from a Global Hate Speech Dataset Representative of a Day on Twitter

Manuel Tonneau, Diyi Liu, Niyati Malhotra, Scott A. Hale, Samuel Fraiberger, Victor Orozco-Olvera, Paul Röttger

Manuel Tonneau, Diyi Liu, Niyati Malhotra, Scott A. Hale, Samuel Fraiberger, Victor Orozco-Olvera, Paul Röttger

14:22 - 14:33

Past Meets Present: Creating Historical Analogy with Large Language Models

Nianqi Li, Siyu Yuan, Jiangjie Chen, Jiaqing Liang, Feng Wei, Zujie Liang, Deqing Yang, Yanghua Xiao

Nianqi Li, Siyu Yuan, Jiangjie Chen, Jiaqing Liang, Feng Wei, Zujie Liang, Deqing Yang, Yanghua Xiao

14:33 - 14:44

What is Stigma Attributed to? A Theory-Grounded, Expert-Annotated Interview Corpus for Demystifying Mental-Health Stigma

Han Meng, Yancan Chen, Yunan Li, YITIAN YANG, Jungup Lee, Renwen Zhang, Yi-Chieh Lee

Han Meng, Yancan Chen, Yunan Li, YITIAN YANG, Jungup Lee, Renwen Zhang, Yi-Chieh Lee

14:44 - 14:55

From Real to Synthetic: Synthesizing Millions of Diversified and Complicated User Instructions with Attributed Grounding

Chiwei Zhu, Benfeng Xu, Xiaorui Wang, Zhendong Mao

Chiwei Zhu, Benfeng Xu, Xiaorui Wang, Zhendong Mao

14:55 - 15:06

Can Large Language Models Detect Errors in Long Chain-of-Thought Reasoning?

Yancheng He, Shilong Li, Jiaheng Liu, Weixun Wang, Xingyuan Bu, Ge Zhang, Z.Y. Peng, Zhaoxiang Zhang, Zhicheng Zheng, Wenbo Su, Bo Zheng

Yancheng He, Shilong Li, Jiaheng Liu, Weixun Wang, Xingyuan Bu, Ge Zhang, Z.Y. Peng, Zhaoxiang Zhang, Zhicheng Zheng, Wenbo Su, Bo Zheng

15:06 - 15:17

AUTALIC: A Dataset for Anti-AUTistic Ableist Language In Context

Naba Rizvi, Harper Strickland, Daniel Gitelman, Tristan Cooper, Alexis Morales Flores, Aekta Kallepalli, Akshat Alurkar, Haaset Owens, Saleha Ahmedi, Isha Khirwadkar, Imani N. S. Munyaka, Nedjma Ousidhoum

Naba Rizvi, Harper Strickland, Daniel Gitelman, Tristan Cooper, Alexis Morales Flores, Aekta Kallepalli, Akshat Alurkar, Haaset Owens, Saleha Ahmedi, Isha Khirwadkar, Imani N. S. Munyaka, Nedjma Ousidhoum

15:17 - 15:28

PARME: Parallel Corpora for Low-Resourced Middle Eastern Languages

Sina Ahmadi, Rico Sennrich, Erfan Karami, Ako Marani, Parviz Fekrazad, Gholamreza Akbarzadeh Baghban, Hanah Hadi, Semko Heidari, Mahîr Dogan, Pedram Asadi, Dashne Bashir, Mohammad Amin Ghodrati, Kourosh Amini, Zeynab Ashourinezhad, Mana Baladi, Farshid Ezzati, Alireza Ghasemifar, Daryoush Hosseinpour, Behrooz Abbaszadeh, Amin Hassanpour, Bahaddin jalal hamaamin, Saya Kamal Hama, Ardeshir Mousavi, Sarko Nazir Hussein, Isar Nejadgholi, Mehmet Ölmez, Horam Osmanpour, Rashid Roshan Ramezani, Aryan Sediq Aziz, Ali Salehi, Mohammadreza Yadegari, Kewyar Yadegari, Sedighe Zamani Roodsari

Sina Ahmadi, Rico Sennrich, Erfan Karami, Ako Marani, Parviz Fekrazad, Gholamreza Akbarzadeh Baghban, Hanah Hadi, Semko Heidari, Mahîr Dogan, Pedram Asadi, Dashne Bashir, Mohammad Amin Ghodrati, Kourosh Amini, Zeynab Ashourinezhad, Mana Baladi, Farshid Ezzati, Alireza Ghasemifar, Daryoush Hosseinpour, Behrooz Abbaszadeh, Amin Hassanpour, Bahaddin jalal hamaamin, Saya Kamal Hama, Ardeshir Mousavi, Sarko Nazir Hussein, Isar Nejadgholi, Mehmet Ölmez, Horam Osmanpour, Rashid Roshan Ramezani, Aryan Sediq Aziz, Ali Salehi, Mohammadreza Yadegari, Kewyar Yadegari, Sedighe Zamani Roodsari

Session 3: Semantics: Lexical and Sentence-Level (Monday 28 July, 14:00-15:30, Room 1.86)

14:00 - 14:11

Token Prepending: A Training-Free Approach for Eliciting Better Sentence Embeddings from LLMs

Yuchen Fu, Zifeng Cheng, Zhiwei Jiang, Zhonghui Wang, Yafeng Yin, Zhengliang Li, Qing Gu

Yuchen Fu, Zifeng Cheng, Zhiwei Jiang, Zhonghui Wang, Yafeng Yin, Zhengliang Li, Qing Gu

14:11 - 14:22

A Text is Worth Several Tokens: Text Embedding from LLMs Secretly Aligns Well with The Key Tokens

Zhijie Nie, Richong Zhang, Zhanyu Wu

Zhijie Nie, Richong Zhang, Zhanyu Wu

14:22 - 14:33

Conditional Dichotomy Quantification via Geometric Embedding

Shaobo Cui, Wenqing Liu, Yiyang Feng, Jiawei Zhou, Boi Faltings

Shaobo Cui, Wenqing Liu, Yiyang Feng, Jiawei Zhou, Boi Faltings

14:33 - 14:44

Modeling the Evolution of English Noun Compounds with Feature-Rich Diachronic Compositionality Prediction

Filip Miletić, Sabine Schulte im Walde

Filip Miletić, Sabine Schulte im Walde

14:44 - 14:55

Modelling Complex Semantics Relation with Contrastively Fine-Tuned Relational Encoders

Naïm Es-sebbani, Esteban Marquer, Zied Bouraoui

Naïm Es-sebbani, Esteban Marquer, Zied Bouraoui

14:55 - 15:06

Learning to Look at the Other Side: A Probing Study of Word Semantics in LLMs with Enabled Bidirectional Attention

Zhaoxin Feng, MA Jianfei, Xiaoyi Bao, Xiaojing Zhao, Emmanuele Chersoni

Zhaoxin Feng, MA Jianfei, Xiaoyi Bao, Xiaojing Zhao, Emmanuele Chersoni

15:06 - 15:17

MEXMA: Token-level objectives improve sentence representations

João Maria Janeiro, Benjamin Piwowarski, Patrick Gallinari, Loic Barrault

João Maria Janeiro, Benjamin Piwowarski, Patrick Gallinari, Loic Barrault

15:17 - 15:28

Representations of Fact, Fiction and Forecast in Large Language Models: Epistemics and Attitudes

Meng Li, Michael Vrazitulis, David Schlangen

Meng Li, Michael Vrazitulis, David Schlangen

Session 3: Special Theme: Generalization of NLP Models (Monday 28 July, 14:00-15:30, Room 1.15-16)

14:00 - 14:11

On the Reliability of Large Language Models for Causal Discovery
Tao Feng, Lizhen Qu, Niket Tandon, Zhuang Li, Xiaoxi Kang, Gholamreza Haffari
Tao Feng, Lizhen Qu, Niket Tandon, Zhuang Li, Xiaoxi Kang, Gholamreza Haffari

14:11 - 14:22

Cross-lingual Generalization and Compression: From Language-Specific to Shared Neurons
Frederick Riemenschneider, Anette Frank
Frederick Riemenschneider, Anette Frank

14:22 - 14:33

Lexical Recall or Logical Reasoning: Probing the Limits of Reasoning Abilities in Large Language Models
Henrike Beyer, Chris Reed
Henrike Beyer, Chris Reed

14:33 - 14:44

Linguistic Generalizability of Test-Time Scaling in Mathematical Reasoning
Guijin Son, Jiwoo Hong, Hyunwoo Ko, James Thorne
Guijin Son, Jiwoo Hong, Hyunwoo Ko, James Thorne

14:44 - 14:55

Limited Generalizability in Argument Mining: State-Of-The-Art Models Learn Datasets, Not Arguments
Marc Feger, Katarina Boland, Stefan Dietze
Marc Feger, Katarina Boland, Stefan Dietze

14:55 - 15:06

M-IFEval: On Multilingual Instruction-Following Capability of Large Language Models
Bo Zeng, Chenyang Lyu, Sinuo Liu, Mingyan Zeng, Minghao Wu, Xuanfan Ni, Tianqi Shi, Yu Zhao, Yefeng Liu, Chenyu Zhu, Ruizhe Li, Jiahui Geng, Qing Li, Yu Tong, Longyue Wang, Weihua Luo, Kaifu Zhang
Bo Zeng, Chenyang Lyu, Sinuo Liu, Mingyan Zeng, Minghao Wu, Xuanfan Ni, Tianqi Shi, Yu Zhao, Yefeng Liu, Chenyu Zhu, Ruizhe Li, Jiahui Geng, Qing Li, Yu Tong, Longyue Wang, Weihua Luo, Kaifu Zhang

15:06 - 15:17

TiC-LM: A Web-Scale Benchmark for Time-Continual LLM Pretraining
Jeffrey Li, Mohammadreza Armandpour, Seyed Iman Mirzadeh, Sachin Mehta, Vaishaal Shankar, Raviteja Vemulapalli, Samy Bengio, Oncel Tuzel, Mehrdad Farajtabar, Hadi Pouransari, Fartash Faghri
Jeffrey Li, Mohammadreza Armandpour, Seyed Iman Mirzadeh, Sachin Mehta, Vaishaal Shankar, Raviteja Vemulapalli, Samy Bengio, Oncel Tuzel, Mehrdad Farajtabar, Hadi Pouransari, Fartash Faghri

Session 3: Student Research Workshop (Monday 28 July, 14:00-15:30, Hall N.2)

14:00 - 14:10

Reasoning for Translation: Comparative Analysis of Chain-of-Thought and Tree-of-Thought Prompting for LLM Translation

Lam Nguyen, Yang Xu

As Large Language Models (LLMs) continue to advance in capability, prompt engineering has emerged as a crucial method for optimizing their performance on specialized tasks. While prompting strategies like Zero-shot, Few-shot, Chain-of-Thought, and Tree-of-Thought have demonstrated significant improvements in reasoning tasks, their application to machine translation has received comparatively less attention. This paper systematically evaluates these prompting techniques across diverse language pairs and domains, measuring their effect on translation quality. Our findings reveal substantial performance variations between prompting methods, with certain strategies offering consistent improvements for specific language directions and complexity levels. These results provide valuable insights for developing more effective LLM-based translation systems without requiring model fine-tuning and complement existing works in the field.

14:00 - 14:10

Towards Multi-Perspective NLP Systems: A Thesis Proposal

Benedetta Muscato

In the field of Natural Language Processing (NLP), a common approach for resolving human disagreement involves establishing a consensus among multiple annotators. However, previous research shows that overlooking individual opinions can result in the marginalization of minority perspectives, particularly in subjective tasks, where annotators may systematically disagree due to their personal preferences. Emerging \textit{Multi-Perspective} approaches challenge traditional methodologies that treat disagreement as mere noise, instead recognizing it as a valuable source of knowledge shaped by annotators' diverse backgrounds, life experiences, and values. This thesis proposal aims to (1) identify the challenges of designing disaggregated datasets i.e., preserving individual labels in human-annotated datasets for subjective tasks (2) propose solutions for developing Perspective-Aware by design systems and (3) explore the correlation between human disagreement and model uncertainty leveraging eXplainable AI techniques (XAI). Our long-term goal is to create a framework adaptable to various subjective NLP tasks to promote the development of more responsible and inclusive models.

14:10 - 14:20

Fact from Fiction: Finding Serialized Novels in Newspapers

Pascale Feldkamp, Allie Lassche, Katrine Frkjr Baunvig, Kristoffer Nielbo, Yuri Bizzoni

Digitized literary corpora of the 19th century favor canonical and novelistic forms, sidelining a broader and more diverse literary production. Serialized fiction – widely read but embedded in newspapers – remains especially underexplored, particularly in low-resource languages like Danish. This paper addresses this gap by developing methods to identify fiction in digitized Danish newspapers (1818–1848). We (1) introduce a manually annotated dataset of 1,394 articles and (2) evaluate classification pipelines using both selected linguistic features and embeddings, achieving F1-scores of up to 0.91. Finally, we (3) analyze feuilleton fiction via interpretable features to test its drift in discourse from neighboring nonfiction. Our results support the construction of alternative literary corpora and contribute to ongoing work on modeling the fiction–nonfiction boundary by operationalizing discourse-level distinctions at scale.

14:20 - 14:30

Semantic alignment in hyperbolic space for fine-grained emotion classification

Ashish Kumar, Durga Toshniwal

Existing approaches to fine-grained emotion classification (FEC) often operate in Euclidean space, where the flat geometry limits the ability to distinguish semantically similar emotion labels (e.g., *annoyed* vs. *angry*). While prior research has explored hyperbolic geometry to capture fine-grained label distinctions, it typically relies on predefined hierarchies and overlooks semantically confusable negatives. In this work, we propose HyCoEM, a semantic alignment framework that leverages the Lorentz model of hyperbolic space. Our approach jointly embeds text and label representations into hyperbolic space via the exponential map, and employs a contrastive loss to bring text embeddings closer to their true labels while pushing them away from adaptively selected, semantically similar negatives. This enables the model to learn label embeddings without relying on a predefined hierarchy and better captures subtle distinctions by incorporating information from both positive and challenging negative labels. Experimental results on two benchmark FEC datasets demonstrate the effectiveness of our approach over baseline methods.

14:30 - 14:40

SimBench: Benchmarking the Ability of Large Language Models to Simulate Human Behaviors

Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, Paul Rttger

Simulations of human behavior based on large language models (LLMs) have the potential to revolutionize the social and behavioral sciences, if and only if they faithfully reflect real human behaviors. Prior work across many disciplines has evaluated the simulation capabilities of specific LLMs in specific experimental settings, but often produced disparate results. To move towards a more robust understanding, we introduce SimBench, the first large-scale benchmark to evaluate how well LLMs can simulate group-level human behaviors across diverse settings and tasks. SimBench compiles 20 datasets in a unified format, measuring diverse types of behavior (e.g., decision-making vs. self-assessment) across hundreds of thousands of diverse participants from different parts of the world. Using SimBench, we can ask fundamental questions regarding when, how, and why LLM simulations succeed or fail. For example, we show that, while even the best LLMs today have limited simulation ability, there is a clear log-linear scaling relationship with model size, and a strong correlation between simulation and scientific reasoning abilities. We also show that base LLMs, on average, are better at simulating high-entropy response distributions, while the opposite holds for instruction-tuned LLMs. By making progress measurable, we hope that SimBench can accelerate the development of better LLM simulators in the future.

14:40 - 14:50

Adversarial Tokenization

Renato Geh, Zilei Shao, Guy Van den Broeck

Current LLM pipelines account for only one possible tokenization for a given string, ignoring exponentially many alternative tokenizations during training and inference. For example, the standard Llama3 tokenization of penguin is [p,enguin], yet [peng,uin] is another perfectly valid alternative. In this paper, we show that despite LLMs being trained solely on one tokenization, they still retain semantic understanding of other tokenizations, raising questions about their implications in LLM safety. Put succinctly, we answer the following question: can we adversarially tokenize an obviously malicious string to evade safety and alignment restrictions? We show that not only is adversarial tokenization an effective yet previously neglected axis of attack, but it is also competitive against existing state-of-the-art adversarial approaches without changing the text of the harmful request. We empirically validate this exploit across three state-of-the-art LLMs and adversarial datasets, revealing a previously unknown vulnerability in subword models.

14:50 - 15:00

Tree-of-Report: Table-to-Text Generation for Sports Game Reports with Tree-Structured Prompting

Shang-Hsuan Chiang, Tsan-Tsung Yang, Kuang-Da Wang, Wei-Yao Wang, An-Zi Yen, Wen-Chih Peng
Generating sports game reports from structured table data is a challenging table-to-text generation task that requires balancing structured data comprehension with narrative storytelling. While model-based approaches demand large training datasets, prompt-based methods with large language models (LLMs) often suffer from hallucination issues due to poor table comprehension. To address these challenges, we propose Tree-of-Report, a novel framework that divides the generation process into three stages: Content Planning, Operation Execution, and Content Generating. Our method decomposes large tables into smaller sub-tables using a hierarchical tree structure, enabling more effective table comprehension. Additionally, it merges and rewrites texts to produce more detailed and coherent long-form outputs. Experimental results on the RotoWire, MLB, and ShuttleSet+ datasets show that Tree-of-Report outperforms existing prompt-based baselines with relatively lower time and cost, demonstrating its advantage in both effectiveness and efficiency. In summary, this work sets a new precedent for prompt-based table-to-text generation in sports game reports.

15:00 - 15:10

NovelHopQA: Diagnosing Multi-Hop Reasoning Failures in Long Narrative Contexts

Abhay Gupta, Michael Lu, Kevin Zhu, Sean O'Brien, Vasu Sharma
Current large language models (LLMs) struggle to answer questions that span tens of thousands of tokens, especially when multi-hop reasoning is involved. While prior benchmarks explore long-context comprehension or multi-hop reasoning in isolation, none jointly vary context length and reasoning depth in natural narrative settings. We introduce NovelHopQA, the first benchmark to evaluate 1–4 hop QA over 64k–128k-token excerpts from 83 full-length public-domain novels. A keyword-guided pipeline builds hop-separated chains grounded in coherent storylines. We evaluate six state-of-the-art (SOTA) models and apply golden context filtering to ensure all questions are genuinely answerable. Human annotators validate both alignment and hop depth. We noticed consistent accuracy drops with increased hops and context length, even in frontier models—revealing that sheer scale does not guarantee robust reasoning. Our failure mode analysis highlights common breakdowns, such as missed final-hop integration and long-range drift. NovelHopQA offers a controlled diagnostic setting to stress-test multi-hop reasoning at scale.

Session 3: TACL Journal (Monday 28 July, 14:00-15:30, Hall M.2)

15:17 - 15:28

NLP Security and Ethics, in the Wild

Heather Lent Erick Galinkin Yiyi Chen Jens Myrup Pedersen Leon Derczynski Johannes Bjerva
As NLP models are used by a growing number of end-users, an area of increasing importance is NLP Security (NLPSec): assessing the vulnerability of models to malicious attacks and developing comprehensive countermeasures against them. While work at the intersection of NLP and cybersecurity has the potential to create safer NLP for all, accidental oversights can result in tangible harm (e.g., breaches of privacy or proliferation of malicious models). In this emerging field, however, the research ethics of NLP have not yet faced many of the long-standing conundrums pertinent to cybersecurity, until now. We thus examine contemporary works across NLPSec, and explore their engagement with cybersecurity's ethical norms. We identify trends across the literature, ultimately finding alarming gaps on topics like harm minimization and responsible disclosure. To alleviate these concerns, we provide concrete recommendations to help NLP researchers navigate this space more ethically, bridging the gap between traditional cybersecurity and NLP ethics, which we frame as ``white hat NLP''. The goal of this work is to help cultivate an intentional culture of ethical research for those working in NLP Security.

Session 3: TACL Journal (Monday 28 July, 14:00-15:30, Room 1.15-16)

15:17 - 15:28

Self-Rationalization in the Wild: A Large Scale Out-of-Distribution Evaluation on NLI-related tasks

Jing Yang Max Glockner Anderson Rocha Iryna Gurevych

Free-text explanations are expressive and easy to understand, but many datasets lack annotated explanation data, making it challenging to train models for explainable predictions. To address this, we investigate how to use existing explanation datasets for self-rationalization and evaluate models' out-of-distribution (OOD) performance. We fine-tune T5-Large and OLMo-7B models and assess the impact of fine-tuning data quality, the number of fine-tuning samples, and few-shot selection methods. The models are evaluated on 19 diverse OOD datasets across three tasks: natural language inference, fact-checking, and hallucination detection in abstractive summarization. For the generated explanation evaluation, we conduct a human study on 13 selected models and study its correlation with the Acceptability score (T5-11B) and three other LLM-based reference-free metrics. Human evaluation shows that the Acceptability score correlates most strongly with human judgments, demonstrating its effectiveness in evaluating free-text explanations. Our findings reveal: 1) few annotated examples effectively adapt models for OOD explanation generation; 2) compared to sample selection strategies, fine-tuning data source has a larger impact on OOD performance; and 3) models with higher label prediction accuracy tend to produce better explanations, as reflected by higher Acceptability scores.

Session 3: TACL Journal (Monday 28 July, 14:00-15:30, Room1.85)

14:33 - 14:44

A Context-aware Framework for Translation-mediated Conversations

José Pombal Sweta Agrawal Emmanouil Zaranis Patrick Fernandes André Martins

Automatic translation systems offer a powerful solution to bridge language barriers in scenarios where participants do not share a common language. However, these systems can introduce errors leading to misunderstandings and conversation breakdown. A key issue is that current systems fail to incorporate the rich contextual information necessary to resolve ambiguities and omitted details, resulting in literal, inappropriate, or misaligned translations. In this work, we present a framework to improve large language model-based translation systems by incorporating contextual information in bilingual conversational settings during training and inference. We validate our proposed framework on two task-oriented domains: customer chat and user-assistant interaction. Across both settings, our framework consistently results in better translations than state-of-the-art systems like GPT-4o and TowerInstruct, as measured by multiple automatic translation quality metrics on several language pairs. We also show that the resulting model leverages context in an intended and interpretable way, improving consistency between the conveyed message and the generated translations.

Poster Presentation

Session 2: CL Journal (Monday 28 July, 11:00-12:30, Hall 4/5)

Relation Extraction in underexplored biomedical domains: A diversity-optimised sampling and synthetic data generation approach

Maxime Delmas, Magdalena Wysocka, André Freitas

The sparsity of labelled data is an obstacle to the development of Relation Extraction (RE) models and the completion of databases in various biomedical areas. While being of high interest in drug-discovery, the literature on natural-products, reporting the identification of potential bioactive compounds from organisms, is a concrete example of such an overlooked topic. To mark the start of this new task, we created the first curated evaluation dataset and extracted literature items from the LOTUS database to build training sets. To this end, we developed a new sampler, inspired by diversity metrics in ecology, named Greedy Maximum Entropy sampler (https://github.com/idiap/gme-sampler). The strategic optimization of both balance and diversity of the selected items in the evaluation set is important given the resource-intensive nature of manual curation. After quantifying the noise in the training set, in the form of discrepancies between the text of input abstracts and the expected output labels, we explored different strategies accordingly. Framing the task as an end-to-end Relation Extraction, we evaluated the performance of standard fine-tuning (BioGPT, GPT-2 and Seq2rel) and few-shot learning with open Large Language Models (LLaMA 7B-65B). In addition to their evaluation in few-shot settings, we explore the potential of open LLMs as synthetic data generators and propose a new workflow for this purpose. All evaluated models exhibited substantial improvements when fine-tuned on synthetic abstracts rather than the original noisy data. We provide our best performing (f1-score=59.0) BioGPT-Large model for end-to-end RE of natural-products relationships along with all the training and evaluation datasets. See more details at https://github.com/idiap/abroad-re.

Compositionality and Sentence Meaning: Comparing Semantic Parsing and Transformers on a Challenging Sentence Similarity Dataset

James Bradley Fodor, Simon De Deyne, Shinsuke Suzuki

One of the major outstanding questions in computational semantics is how humans integrate the meaning of individual words into a sentence in a way that enables understanding of complex and novel combinations of words, a phenomenon known as compositionality. Two major approaches to modelling the process of compositionality are vector-based models, in which the meaning of a sentence is represented as a vector of numbers, and syntax-based models, in which the meaning of a sentence is represented as a structured tree of labelled components. A major barrier in assessing and comparing these contrasting approaches is the lack of large, relevant datasets for model comparison. This paper aims to address this gap by introducing a new dataset, STS3k, which consists of 2,800 pairs of sentences rated for semantic similarity by human participants. The sentence pairs have been selected to systematically vary different combinations of words, providing a rigorous test and enabling a clearer picture of the comparative strengths and weaknesses of vector-based and syntax-based methods. Our results show that when tested on the new STS3k dataset, state-of-the-art transformers poorly capture the pattern of human semantic similarity judgments, while even simple methods for combining syntax- and vector-based components into a novel hybrid model yield substantial improvements. We further show that this improvement is due to the ability of the hybrid model to replicate human sensitivity to specific changes in sentence structure. Our findings provide evidence for the value of integrating multiple methods to better reflect the way in which humans mentally represent compositional meaning.

Exploring temporal sensitivity in the brain using multi-timescale language models: an EEG decoding study

Sijie Ling, Alex Murphy, Alona Fyshe

The brain's ability to perform complex computations at varying timescales is crucial, ranging from understanding single words to grasping the overarching narrative of a story. Recently, multi-timescale long short-term memory (MT-LSTM) models (Mahto et al. 2020; Jain et al. 2020) have been introduced to use temporally-tuned parameters to induce sensitivity to different timescales of language processing (i.e. related to near/distant words). However, whether the temporally-tuned information processing in MT-LSTMs is similar to the brain has not been explored using high temporal resolution recording modalities such as electroencephalography (EEG).To bridge this gap, we used an EEG dataset recorded while participants listened to Chapter 1 of "Alice in Wonderland" and trained ridge regression models to predict the temporally-tuned MT-LSTM embeddings from EEG responses. Our analysis reveals that EEG signal segments can effectively predict MT-LSTM embeddings across various timescales. For longer timescales, predictions are significant within an extended time window of Â±2 s around word onset, while for shorter timescales, significant predictions are confined to a narrow window ranging from -180 ms to 790 ms. Intriguingly, we observed that short timescale information is not only processed in the vicinity of word onset but also at distant time points. % , specifically at -1 and +2 s.These observations underscore the parallels and discrepancies between computational models and the neural mechanisms of the brain. As word embeddings are used more as \textit{in silico} models of semantic representation in the brain, the ability to induce word embeddings, tuned to different timescales, allows for an ever-increasing range of questions in language processing in humans and machines to be addressed via brain-model analyses.

Usage-based Grammar Induction from Minimal Cognitive Principles

Anna Jon-And, Jérôme Michaud

This study explores the cognitive mechanisms underlying human language acquisition through grammar induction by a minimal cognitive architecture, with a short and flexible sequence memory as its most central feature. We use reinforcement learning for the task of identifying sentences in a stream of words from artificial languages. Results demonstrate the modelâ_x0080__x0099_s ability to identify frequent and informative multi-word chunks, reproducing characteristics of natural language acquisition. The model successfully navigates varying degrees of linguistic complexity, exposing efficient adaptation to combinatorial challenges through the reuse of sequential patterns. The emergence of parsimonious tree structures suggests an optimization for the sentence identification task, balancing economy and information. The cognitive architecture reflects aspects of human memory systems and decision-making processes, enhancing its cognitive plausibility. While the model exhibits limitations in generalization and semantic representation, its minimalist nature offers insights into some fundamental mechanisms of language learning. Our study demonstrates the power of this simple architecture and stresses the importance of sequence memory in language learning. Since other animals do not seem to have faithful sequence memory, this may be a key to understanding why only humans have developed complex languages.

Humans Learn Language from Situated Communicative Interactions. What about Machines?

Paul Van Eecke, Katrien Beuls

Humans acquire their native languages by actively taking part in communicative interactions with their caregivers. These interactions are meaningful, intentional and situated in their everyday environment. The situated and communicative nature of the interactions is essential to the language acquisition process, as language learners depend on clues provided by the communicative environment to make sense of the utterances they perceive. As such, the linguistic knowledge they build up is rooted in linguistic forms, their meaning and their communicative function. When it comes to machines, the situated, communicative and interactional aspects of language learning are often passed over. This applies in particular to today's large language models (LLMs), where the input is purely text-based, and where the distribution of character groups or words serves as a basis for modelling the meaning of linguistic expressions. In this paper, we argue that this design choice lies at the root of a number of important limitations, in particular regarding the data hungriness of the models, their limited ability to perform logical and pragmatic reasoning, and their susceptibility to biases. At the same time, we make a case for an alternative approach that models how artificial agents can acquire linguistic structures by participating in situated communicative interactions. Through a selection of experiments, we show how the linguistic knowledge that is captured in the resulting models is of a fundamentally different nature than the knowledge captured by LLMs, and argue that this change of perspective provides a promising path towards more human-like language processing in machines.

Neural Semantic Parsing with Extremely Rich Symbolic Meaning Representations

Xiao Zhang, Gosse Bouma, Johan Bos

Current open-domain neural semantics parsers show impressive performance. However, closer inspection of the symbolic meaning representations they produce reveals significant weaknesses: sometimes they tend to merely copy character sequences from the source text to form symbolic concepts, defaulting to the most frequent word sense based in the training distribution. By leveraging the hierarchical structure of a lexical ontology, we introduce a novel compositional symbolic representation for concepts based on their position in the taxonomical hierarchy. This representation provides richer semantic information and enhances interpretability. We introduce a neural "taxonomical" semantic parser Â to utilize this new representation system of predicates, and compare it with a standard neural semantic parser trained on the traditional meaning representation format, employing a novel challenge set and evaluation metric for evaluation. Our experimental findings demonstrate that the taxonomical model, trained on much richer and complex meaning representations, is slightly subordinate in performance to the traditional model using the standard metrics for evaluation, but outperforms it when dealing with out-of-vocabulary concepts. We further showed through neural model probing that training on a taxonomic representation enhances the model's ability to learn the taxonomical hierarchy. This finding is encouraging for research in computational semantics that aims to combine data-driven distributional meanings with knowledge-based symbolic representations.

Investigating Idiomaticity in Word Representations

Aline Villavicencio, Wei He, Marcos Garcia Gonzalex, Tiago K Vieira, Carolina Scarton, Marco Idiart

Idiomatic expressions are an integral part of human languages, often used to express complex ideas in compressed or conventional ways (e.g., eager beaver as a keen and enthusiastic person). However, their interpretations may not be straightforwardly linked to the meanings of their individual components in isolation and this may have an impact for compositional approaches. In this article, we investigate to what extent word representation models are able to go beyond compositional word combinations and capture multiword expression idiomaticity and some of the expected properties related to idiomatic meanings. We focus on noun compounds of varying levels of idiomaticity in two languages (English and Portuguese), presenting a dataset of minimal pairs containing human idiomaticity judgments for each noun compound at both type and token levels, their paraphrases and their occurrences in naturalistic and sense-neutral contexts, totalling 32,200 sentences. We propose this set of minimal pairs for evaluating how well a model captures idiomatic meanings, and define a set of fine-grained metrics of Affinity and Scaled Similarity, to determine how sensitive the models are to perturbations that may lead to changes in idiomaticity. Affinity is a comparative measure of the similarity between an experimental item, a target and a potential distractor, and Scaled Similarity incorporates a rescaling factor to magnify the meaningful similarities within the spaces defined by each specific model. The results obtained with a variety of representative and widely used models indicate that, despite superficial indications to the contrary in the form of high similarities, idiomaticity is not yet accurately represented in current models. Moreover, the performance of models with different levels of contextualization suggests that their ability to capture context is not yet able to go beyond more superficial lexical clues provided by the words and to actually incorporate the relevant semantic clues needed for idiomaticity. By proposing model-agnostic measures for assessing the ability of models to capture idiomaticity, this article contributes to determining limitations in the handling of non-compositional structures, which is one of the directions that needs to be considered for more natural, accurate, and robust language understanding. The source code and additional materials related to this paper are available at our GitHub repository.

Session 2: Dialogue and Interactive Systems (Monday 28 July, 11:00-12:30, Hall 4/5)

Speaking Beyond Language: A Large-Scale Multimodal Dataset for Learning Nonverbal Cues from Video-Grounded Dialogues

Youngmin Kim, Jiwan Chung, Jisoo Kim, sunghyun lee, Sangkyu Lee, Junhyeok Kim, Cheoljong Yang, Youngjae Yu

Nonverbal communication is integral to human interaction, with gestures, facial expressions, and body language conveying critical aspects of intent and emotion. However, existing large language models (LLMs) fail to effectively incorporate these nonverbal elements, limiting their capacity to create fully immersive conversational experiences. We introduce MARS, a multimodal language model designed to understand and generate nonverbal cues alongside text, bridging this gap in conversational AI. Our key innovation is VENUS, a large-scale dataset comprising annotated videos with time-aligned text, facial expressions, and body language. Leveraging VENUS, we train MARS with a next-token prediction objective, combining text with vector-quantized nonverbal representations to achieve multimodal understanding and generation within a unified framework. Based on various analyses of the VENUS datasets, we validate its substantial scale and high effectiveness. Our quantitative and qualitative results demonstrate that MARS successfully generates text and nonverbal languages, corresponding to conversational input. Our dataset and code are available at https://github.com/winston1214/nonverbal-conversation.

X-TURING: Towards an Enhanced and Efficient Turing Test for Long-Term Dialogue Agents

Weiqi Wu, Hongqiu Wu, hai zhao

The Turing test examines whether AIs exhibit human-like behaviour in natural language conversations. The traditional setting limits each participant to one message at a time and requires constant human participation. This fails to reflect a natural conversational style and hinders the evaluation of dialogue agents based on Large Language Models (LLMs) in complex and prolonged interactions. This paper proposes X-Turing, which enhances the original test with a burst dialogue pattern, allowing more dynamic exchanges using consecutive messages. It further reduces human workload by iteratively generating dialogues that simulate the long-term interaction between the agent and a human to compose the majority of the test process. With the pseudo-dialogue history, the agent then engages in a shorter dialogue with a real human, which is paired with a human-human conversation on the same topic to be judged using questionnaires. We introduce the X-Turn Pass-Rate metric to assess the human likeness of LLMs across varying durations. While LLMs like GPT-4 initially perform well, achieving pass rates of 51.9% and 38.9% during 3 turns and 10 turns of dialogues respectively, their performance drops as the dialogue progresses, which underscores the difficulty in maintaining consistency in the long term.

CONFETTI: Conversational Function-Calling Evaluation Through Turn-Level Interactions

Tamer Alkhouli, Katerina Margatina, James Gung, Raphael Shu, Claudia Zaghi, MONICA SUNKARA, Yi Zhang

We introduce Conversational Function-Calling Evaluation Through Turn-Level Interactions (CONFETTI), a conversational benchmark designed to evaluate the function-calling capabilities and response quality of large language models (LLMs). Current benchmarks lack comprehensive assessment of LLMs in complex conversational scenarios. CONFETTI addresses this gap through 109 human-simulated conversations, comprising 313 user turns and covering 86 APIs. These conversations explicitly target various conversational complexities, such as follow-ups, goal correction and switching, ambiguous and implicit goals. We perform off-policy turn-level evaluation using this benchmark targeting function-calling. Our benchmark also incorporates dialog act annotations to assess agent responses. We evaluate a series of state-of-the-art LLMs and analyze their performance with respect to the number of available APIs, conversation lengths, and chained function calling. Our results reveal that while some models are able to handle long conversations, and leverage more than 20+ APIs successfully, other models struggle with longer context or when increasing the number of APIs. We also report that the performance on chained function-calls is severely limited across the models. Overall, the top performing models on CONFETTI are Nova Pro (40.01%), Claude Sonnet v3.5 (35.46%) and Llama 3.1 405B (33.19%) followed by command-r-plus (31.18%) and Mistral-Large-2407 (30.07%).

Exploring the Potential of LLMs as Personalized Assistants: Dataset, Evaluation, and Analysis

Jisoo Mok, Ik-hwan Kim, Sangkwon Park, Sungroh Yoon

Personalized AI assistants, a hallmark of the human-like capabilities of Large Language Models (LLMs), are a challenging application that intertwines multiple problems in LLM research. Despite the growing interest in the development of personalized assistants, the lack of an open-source conversational dataset tailored for personalization remains a significant obstacle for researchers in the field. To address this research gap, we introduce HiCUPID, a new benchmark to probe and unleash the potential of LLMs to deliver personalized responses. Alongside a conversational dataset, HiCUPID provides a Llama-3.2-based automated evaluation model whose assessment closely mirrors human preferences. We release our dataset, evaluation model, and code at https://github.com/12kimih/HiCUPID.

Beware of Your Po! Measuring and Mitigating AI Safety Risks in Role-Play Fine-Tuning of LLMs

Weixiang Zhao, Yulin Hu, Yang Deng, Jiahe Guo, Xingyu Sui, Xinyang Han, An Zhang, Yanyan Zhao, Bing Qin, Tat-Seng Chua, Ting Liu

Role-playing enables large language models (LLMs) to engage users in immersive and personalized interactions, but it also introduces significant safety risks. Existing role-play fine-tuning techniques improve role adaptability but may degrade safety performance, particularly for villainous characters. In this work, we conduct the first comprehensive assessment of role-play fine-tuning risks by training 95 role-specific LLMs using RoleBench. Our experiments reveal that role-play fine-tuning leads to a noticeable decline in safety performance, with safety risks varying based on character traits. To tackle this challenge, we propose Safety-Aware Role-Play Fine-Tuning (SaRFT), a novel method designed to balance role-playing capabilities and safety. Extensive experiments on LLaMA-3-8B-Instruct, Gemma-2-9B-it, and Qwen2.5-7B-Instruct demonstrate that SaRFT consistently outperforms state-of-the-art baselines under both LoRA and full-parameter fine-tuning settings. Our findings highlight the necessity of role-adaptive safety measures and provide insights into mitigating role-specific safety risks in role-playing LLMs.

Beyond Dialogue: A Profile-Dialogue Alignment Framework Towards General Role-Playing Language Model

Yeyong Yu, Runsheng Yu, Haojie Wei, Zhanqiu Zhang, Quan QIAN

The rapid advancement of large language models (LLMs) has revolutionized role-playing, enabling the development of general role-playing models. However, current role-playing training has two significant issues: (I) Using a predefined role profile to prompt dialogue training for specific scenarios usually leads to biases and even conflicts between the dialogue and the profile, resulting in training biases. (II) Models learn to imitate the role based solely on the profile, neglecting profile-dialogue alignment at the sentence level. To overcome the aforementioned hurdles, we propose a novel framework **Beyond Dialogue**, which introduces "beyond dialogue" tasks to align dialogue with profile traits for each scenario, eliminating biases during training. Furthermore, the framework achieves a sentence-level fine-grained alignment between profile and dialogue through an innovative prompting mechanism that generates reasoning data for training. Moreover, the aforementioned methods are fully automated and low-cost. Experimental results demonstrate our model excels in adhering to role profiles, outperforming most proprietary general and specialized role-playing baselines. The code and data are provided in https://github.com/yuyouyu32/BeyondDialogue.

RolePlot: A Systematic Framework for Evaluating and Enhancing the Plot-Progression Capabilities of Role-Playing Agents

Pinyi Zhang, Siyu An, Lingfeng Qiao, Yifei Yu, Jingyang Chen, Jie Wang, di yin, Xing Sun, Kai Zhang

Role-playing agents (RPAs) are garnering increasing interests as a novel form of conversational AI. While previous research has predominantly concentrated on their ability to portray specified characters, we argue from a user-centered perspective that RPAs' capability to advance the plot requires substantial improvements to deliver more engaging interaction. To bridge this gap, we propose RolePlot, a role-playing framework specifically designed to evaluate and enhance the plot-progression capabilities of RPAs. RolePlot begins by constructing a plot-progression dataset extended from human-written literary scripts and specially designed synthetic data, followed by narrative theory-driven manual annotation and automated labeling validated through human verification. We then exploit the over-parameterized embedding space of LLMs to detect a ``trigger subspace'' that identifies dialogue segments catalyzing plot transitions. When user's inputs align with this subspace, we explicitly prompt RPAs to advance the plot. For evaluation, we simulate User-RPA interactions and track both the conversation longevity (measured in dialogue turns before disengagement) and users' arousal levels across different stages. Empirically, our method improves RPAs' capability to time plot developments, and more importantly, yielding a significant increase in conversation turns and sustained higher arousal levels, thereby confirming that users experience more immersive engagements.

SDPO: Segment-Level Direct Preference Optimization for Social Agents

Aobo Kong, Wentao Ma, Shiwan Zhao, Yongbin Li, Yuchuan Wu, Ke Wang, Xiaoqian Liu, Qicheng Li, Yong Qin, Fei Huang

Social agents powered by large language models (LLMs) can simulate human social behaviors but fall short in handling complex social dialogues. Direct Preference Optimization (DPO) has proven effective in aligning LLM behavior with human preferences across various agent tasks. However, standard DPO focuses solely on individual turns, which limits its effectiveness in multi-turn social interactions. Several DPO-based multi-turn alignment methods with session-level data have shown potential in addressing this problem. While these methods consider multiple turns across entire sessions, they are often overly coarse-grained, introducing training noise, and lack robust theoretical support. To resolve these limitations, we propose Segment-Level Direct Preference Optimization (SDPO), which dynamically select key segments within interactions to optimize multi-turn agent behavior. SDPO minimizes training noise and is grounded in a rigorous theoretical framework. Evaluations on the SOTOPIA benchmark demonstrate that SDPO-tuned agents consistently outperform both existing DPO-based methods and proprietary LLMs like GPT-4o, underscoring SDPO's potential to advance the social intelligence of LLM-based agents. We release our code and data at https://anonymous.4open.science/r/SDPO-CE8F.

SHARE: Shared Memory-Aware Open-Domain Long-Term Dialogue Dataset Constructed from Movie Script

Eunwon Kim, Chanho Park, Buru Chang

Shared memories between two individuals strengthen their bond and are crucial for facilitating their ongoing conversations. This study aims to make long-term dialogue more engaging by leveraging these shared memories. To this end, we introduce a new long-term dialogue dataset named SHARE, constructed from movie scripts, which are a rich source of shared memories among various relationships. Our dialogue dataset contains the summaries of persona information and events of two individuals, as explicitly revealed in their conversation, along with implicitly extractable shared memories. We also introduce EPISODE, a long-term dialogue framework based on SHARE that utilizes shared experiences between individuals. Through experiments using SHARE, we demonstrate that shared memories between two individuals make long-term dialogues more engaging and sustainable, and that EPISODE effectively manages shared memories during dialogue. Our dataset and code are available at https://github.com/e1kim/SHARE.

OmniFlatten: An End-to-end GPT Model for Seamless Voice Conversation

Qinglin Zhang, Luyao Cheng, Chong Deng, Qian Chen, Wen Wang, Siqi Zheng, Jiaqing Liu, Hai Yu, Chao-Hong Tan, Zhihao Du, ShiLiang Zhang

Full-duplex spoken dialogue systems significantly surpass traditional turn-based dialogue systems, as they allow simultaneous bidirectional communication, closely mirroring human-human interactions. However, achieving low latency and natural interactions in full-duplex dialogue systems remains a significant challenge, especially considering human conversation dynamics such as interruptions, backchannels, and overlapping speech. In this paper, we introduce a novel End-to-End GPT-based model OmniFlatten for full-duplex conversation, capable of effectively modeling the complex behaviors inherent to natural conversations with low latency. To achieve full-duplex conversation capabilities, we propose a multi-stage post-training scheme that progressively adapts a text large language model (LLM) backbone into a speech-text dialogue LLM, capable of generating text and speech in real time, without modifying the architecture of the backbone LLM. The training process comprises three stages: modality alignment, half-duplex dialogue learning, and full-duplex dialogue learning. In all training stages, we standardize the data using a flattening operation, which enables unifying the training methods and the GPT backbone across different modalities and tasks. Our approach offers a simple modeling technique and a promising research direction for developing efficient and natural end-to-end full-duplex spoken dialogue systems.

Crab: A Novel Configurable Role-Playing LLM with Assessing Benchmark

Kai He, Yucheng Huang, Wenqing Wang, Delong Ran, Dongming Sheng, Junxuan Huang, Qika Lin, Jiaxing Xu, Wenqiang Liu, Mengling Feng

This study introduces Crab, a novel Configurable Role-Playing (RP) LLM with Assessing Benchmark, which consists of Role-Centric Dataset Curation, Persona-Embodying LLM Construction, and Comprehensive Benchmark Creation for RP dialogue generation. Distinct from traditional RP models that employ only several preset roles, Crab enables dynamic configuration of desired roles, thereby enhancing related flexibility and adaptability. To effectively train RP-LLMs, we curated the largest RP training dataset. The dataset provides a detailed role overview for each dialogue, including character profile, conversation scenario, and tagged topic, capturing a broad range of role-based behaviors, emotions, and interactions. We also noticed that current benchmarks lack both proper evaluation standards and methods. Thus, to validate RP-LLMs' effectiveness, we introduced a new benchmark containing an evaluation standard, a test dataset with manual annotations, and a reward model RoleRM designed to automatically assess specific aspects of RP while aligning with human perception. Sufficient experiments reveal that RoleRM significantly outperforms ChatGPT and other evaluation methods in conducting fine-grained evaluations of RP. Also, RP-LLMs powered by Crab demonstrate superior performance across various fine-grained aspects.

Beyond Single Labels: Improving Conversational Recommendation through LLM-Powered Data Augmentation

Haozhe Xu, Xiaohua Wang, Changze Lv, Xiaoqing Zheng

Conversational recommender systems (CRSs) enhance recommendation quality by engaging users in multi-turn dialogues, capturing nuanced preferences through natural language interactions. However, these systems often face the false negative issue, where items that a user might like are incorrectly labeled as negative during training, leading to suboptimal recommendations. Expanding the label set through data augmentation presents an intuitive solution but faces the challenge of balancing two key aspects: ensuring semantic relevance and preserving the collaborative information inherent in CRS datasets. To address these issues, we propose a novel data augmentation framework that first leverages an LLM-based semantic retriever to identify diverse and semantically relevant items, which are then filtered by a relevance scorer to remove noisy candidates. Building on this, we introduce a two-stage training strategy balancing semantic relevance and collaborative information. Extensive experiments on two benchmark datasets and user simulators demonstrate significant and consistent performance improvements across various recommenders, highlighting the effectiveness of our approach in advancing CRS performance.

ACT: Knowledgeable Agents to Design and Perform Complex Tasks

Makoto Nakatsuji, Shuhei Tateishi, Yasuhiro Fujiwara, Ayaka Matsumoto, Narichika Nomoto, Yoshihide Sato

Large language models enhance collaborative task execution in multi-agent systems. Current studies break complex task into manageable tasks, but agents lack understanding of the overall task and how others approach their tasks, hindering synergy and integration. We propose a method called knowledgeable {\bf {\em A}}gents to design and perform {\bf {\em C}}omplex {\bf {\em T}}asks (ACT), where: (1) Agents independently manage their knowledge and tasks while collaboratively design the complex task into a more comprehensible form. In parallel, each agent also acquires knowledge of others, defined as a structured description of how other agents approach their tasks based on the agent's own task resolution. (2) Each agent updates its knowledge and refines its task through interactions with others. By referencing structured knowledge, they effectively integrate their tasks to collaboratively solve the complex task. Three evaluations including creative writing and tool utilization, show that ACT accurately outperforms existing methods in solving complex tasks.

LLMs Can Simulate Standardized Patients via Agent Coevolution

Zhuoyun Du, LujieZheng, Renjun Hu, Yuyang Xu, Xiawei Li, Ying Sun, Wei Chen, Jian Wu, Haolei Cai, Haochao Ying

Training medical personnel using standardized patients (SPs) remains a complex challenge, requiring extensive domain expertise and role-specific practice. Most research on Large Language Model (LLM)-based simulated patients focuses on improving data retrieval accuracy or adjusting prompts through human feedback. However, this focus has overlooked the critical need for patient agents to learn a standardized presentation pattern that transforms data into human-like patient responses through unsupervised simulations. To address this gap, we propose EvoPatient, a novel simulated patient framework in which a patient agent and doctor agents simulate the diagnostic process through multi-turn dialogues, simultaneously gathering experience to improve the quality of both questions and answers, ultimately enabling human doctor training. Extensive experiments on various cases demonstrate that, by providing only overall SP requirements, our framework improves over existing reasoning methods by more than 10% in requirement alignment and better human preference, while achieving an optimal balance of resource consumption after evolving over 200 cases for 10 hours, with excellent generalizability. Our system will be available at https://github.com/ZJUMAI/EvoPatient

ChatSOP: An SOP-Guided MCTS Planning Framework for Controllable LLM Dialogue Agents

Zhigen Li, Jianxiang Peng, Yanmeng Wang, Yong Cao, Tianhao Shen, Minghui Zhang, Linxi Su, Shang Wu, Yihang Wu, YuQian Wang, Ye Wang, Wei Hu, Jianfeng Li, Shaojun Wang, Jing Xiao, Deyi Xiong

Dialogue agents powered by Large Language Models (LLMs) show superior performance in various tasks. Despite the better user understanding and human-like responses, their **lack of controllability** remains a key challenge, often leading to unfocused conversations or task failure. To address this, we introduce Standard Operating Procedure (SOP) to regulate dialogue flow. Specifically, we propose **ChatSOP**, a novel SOP-guided Monte Carlo Tree Search (MCTS) planning framework designed to enhance the controllability of LLM-driven dialogue agents. To enable this, we curate a dataset comprising SOP-annotated multi-scenario dialogues, generated using a semi-automated role-playing system with GPT-4o and validated through strict manual quality control. Additionally, we propose a novel method that integrates Chain of Thought reasoning with supervised fine-tuning for SOP prediction and utilizes SOP-guided Monte Carlo Tree Search for optimal action planning during dialogues. Experimental results demonstrate the effectiveness of our method, such as achieving a 27.95% improvement in action accuracy compared to baseline models based on GPT-3.5 and also showing notable gains for open-source models. Dataset and codes are publicly available.

When Harry Meets Superman: The Role of The Interlocutor in Persona-Based Dialogue Generation

Daniela Occhipinti, Marco Guerini, Malvina Nissim

Endowing dialogue agents with persona information has proven to significantly improve the consistency and diversity of their generations. While much focus has been placed on aligning dialogues with provided personas, the adaptation to the interlocutor's profile remains largely underexplored. In this work, we investigate three key aspects: (1) a model's ability to align responses with both the provided persona and the interlocutor's; (2) its robustness when dealing with familiar versus unfamiliar interlocutors and topics, and (3) the impact of additional fine-tuning on specific persona-based dialogues. We evaluate dialogues generated with diverse speaker pairings and topics, framing the evaluation as an author identification task and employing both LLM-as-a-judge and human evaluations. By systematically masking or disclosing information about interlocutor, we assess its impact on dialogue generation. Results show that access to the interlocutor's persona improves the recognition of the target speaker, while masking it does the opposite. Although models generalise well across topics, they struggle with unfamiliar interlocutors. Finally, we found that in zero-shot settings, LLMs often copy biographical details, facilitating identification but trivialising the task.

Exploring Persona Sentiment Sensitivity in Personalized Dialogue Generation

Yonghyun Jun, Hwanhee Lee

Personalized dialogue systems have advanced considerably with the integration of user-specific personas into large language models (LLMs). However, while LLMs can effectively generate personalized responses, the influence of persona sentiment on dialogue quality remains underexplored. In this work, we conduct a large-scale analysis of dialogues generated using a range of polarized user profiles. Our experiments reveal that dialogues involving negatively polarized users tend to overemphasize persona attributes. In contrast, positively polarized profiles yield dialogues that selectively incorporate persona information, resulting in smoother interactions. Furthermore, we find that personas with weak or neutral sentiment generally produce lower-quality dialogues. Motivated by these findings, we propose a dialogue generation approach that explicitly accounts for persona polarity by combining a turn-based generation strategy with a profile ordering mechanism and sentiment-aware prompting. Our study provides new insights into the sensitivity of LLMs to persona sentiment and offers guidance for developing more robust and nuanced personalized dialogue systems.

SELF-PERCEPT: Introspection Improves Large Language Models' Detection of Multi-Person Mental Manipulation in Conversations

Danush Khanna, Pratinav Seth, Sidhaarth Sredharan Murali, Aditya Kumar Guru, Siddharth Shukla, Tanuj Tyagi, SANDEEP CHAURASIA, Kripabandhu Ghosh

Mental manipulation is a subtle yet pervasive form of abuse in interpersonal communication, making its detection critical for safeguarding potential victims. However, due to manipulation's nuanced and context-specific nature, identifying manipulative language in complex, multi-turn, and multi-person conversations remains a significant challenge for large language models (LLMs). To address this gap, we introduce the MultiManip dataset, comprising 220 multi-turn, multi-person dialogues balanced between manipulative and non-manipulative interactions, all drawn from reality shows that mimic real-world scenarios. For manipulative interactions, it includes 11 distinct manipulations depicting real-life scenarios. We conduct extensive evaluations of state-of-the-art LLMs, such as GPT-4o and Llama-3.1-8B, employing various prompting strategies. Despite their capabilities, these models often struggle to detect manipulation effectively. To overcome this limitation, we propose SELF-PERCEPT, a novel, two-stage prompting framework inspired by Self-Perception Theory, demonstrating strong performance in detecting multi-person, multi-turn mental manipulation. Our code and data are publicly available at https://github.com/danushkhanna/self-percept .

Open-World Planning via Lifted Regression with LLM-Inferred Affordances for Embodied Agents

Xiaotian Liu, Ali Pesaranghader, HANZE LI, Punyaphat Sukcharoenchaikul, Jaehong Kim, Tanmana Sadhu, Hyejeong Jeon, Scott Sanner

Open-world planning with incomplete knowledge is crucial for real-world embodied AI tasks. Despite that, existing LLM-based planners struggle with long chains of sequential reasoning, while symbolic planners face combinatorial explosion of states and actions for complex domains due to reliance on grounding. To address these deficiencies, we introduce LLM-Regress, an open-world planning approach integrating lifted regression with LLM-generated affordances. LLM-Regress generates sound and complete plans in a compact lifted form, avoiding exhaustive enumeration of irrelevant states and actions. Additionally, it makes efficient use of LLMs to infer goal-related objects and affordances without the need to predefine all possible objects and affordances. We conduct extensive experiments on three benchmarks and show that LLM-Regress significantly outperforms state-of-the-art LLM planners and a grounded planner using LLM-generated affordances.

Know You First and Be You Better: Modeling Human-Like User Simulators via Implicit Profiles

Kuang Wang, Xianfei Li, Shenghao Yang, Li Zhou, Feng Jiang, Haizhou Li

User simulators are crucial for replicating human interactions with dialogue systems, supporting both collaborative training and automatic evaluation, especially for large language models (LLMs). However, current role-playing methods face challenges such as a lack of utterance-level authenticity and user-level diversity, often hindered by role confusion and dependence on predefined profiles of well-known figures. In contrast, direct simulation focuses solely on text, neglecting implicit user traits like personality and conversation-level consistency. To address these issues, we introduce the User Simulator with Implicit Profiles (USP), a framework that infers implicit user profiles from human-machine interactions to simulate personalized and realistic dialogues. We first develop an LLM-driven extractor with a comprehensive profile schema, then refine the simulation using conditional supervised fine-tuning and reinforcement learning with cycle consistency, optimizing at both the utterance and conversation levels. Finally, a diverse profile sampler captures the distribution of real-world user profiles. Experimental results show that USP outperforms strong baselines in terms of authenticity and diversity while maintaining comparable consistency. Additionally, using USP to evaluate LLM on dynamic multi-turn aligns well with mainstream benchmarks, demonstrating its effectiveness in real-world applications.

Enhancing Goal-oriented Proactive Dialogue Systems via Consistency Reflection and Correction

Didi Zhang, Yaxin Fan, Peifeng Li, Qiaoming Zhu

Goal-oriented proactive dialogue systems are designed to guide user conversations seamlessly towards specific objectives by planning a goal-oriented path. However, previous research has focused predominantly on optimizing these paths while neglecting the inconsistencies that may arise between generated responses and dialogue contexts, including user profiles, dialogue history, domain knowledge, and subgoals. To address this issue, we introduce a model-agnostic two-stage Consistency Reflection and Correction (CRC) framework. Specifically, in the consistency reflection stage, the model is prompted to reflect on the discrepancies between generated responses and dialogue contexts, identifying inconsistencies and suggesting possible corrections. In the consistency correction stage, the model generates responses that are more consistent with the dialogue context based on these reflection results. We conducted experiments on various model architectures with different parameter sizes, including encoder-decoder models (BART, T5) and decoder-only models (GPT-2, DialoGPT, Phi3, Mistral and LLaMA3), and the experimental results on three datasets demonstrate that our CRC framework significantly improves the consistency between generated responses and dialogue contexts.

Caution for the Environment: LLM Agents are Susceptible to Environmental Distractions

Xinbei Ma, Yiting Wang, Yao Yao, Tongxin Yuan, Aston Zhang, Zhuosheng Zhang, hai zhao

This paper investigates the faithfulness of multimodal large language model (MLLM) agents in a graphical user interface (GUI) environment, aiming to address the research question of whether multimodal GUI agents can be distracted by environmental context. A general scenario is proposed where both the user and the agent are benign, and the environment, while not malicious, contains unrelated content. A wide range of MLLMs are evaluated as GUI agents using a simulated dataset, following three working patterns with different levels of perception. Experimental results reveal that even the most powerful models, whether generalist agents or specialist GUI agents, are susceptible to distractions. While recent studies predominantly focus on the helpfulness of agents, our findings first indicate that these agents are prone to environmental distractions. Furthermore, we implement an adversarial environment injection and analyze the approach to improve faithfulness, calling for a collective focus on this important topic.

Advancing Collaborative Debates with Role Differentiation through Multi-Agent Reinforcement Learning

Haoran Li, Ziyi Su, Yun Xue, Zhiliang Tian, YIPING SONG, Minlie Huang

Multi-agent collaborative tasks exhibit exceptional capabilities in natural language applications and generation. By prompting agents to assign clear roles, it is possible to facilitate cooperation and achieve complementary capabilities among LLMs. A common strategy involves adopting a relatively general role assignment mechanism, such as introducing a ``judge'' or a ``summarizer''. However, these approaches lack task-specific role customization based on task characteristics. Another strategy involves decomposing the task based on domain knowledge and task characteristics, followed by assigning appropriate roles according to LLMs' respective strengths, such as programmers and testers. However, in some given tasks, obtaining domain knowledge related to task characteristics and getting the strengths of different LLMs is hard. To solve these problems, we propose a Multi-LLM Cooperation (MLC) framework with automatic role assignment capabilities. The core idea of the MLC is to initialize role assignments randomly and then allow the role embeddings to be learned jointly with the downstream task. To capture the state transitions of multiple LLMs during turn-based speaking, the role embedding is sequence-aware. At the same time, to avoid role convergence, the role differentiation module in MLC encourages behavioral differentiation between LLMs while ensuring the LLM team consistency, guiding different LLMs to develop complementary strengths from the optimization level. Our experiments on seven datasets demonstrate that MLC significantly enhances collaboration and expertise, which collaboratively addresses multi-agent tasks.

AgentDropout: Dynamic Agent Elimination for Token-Efficient and High-Performance LLM-Based Multi-Agent Collaboration

Zhexuan Wang, Yutong Wang, Xuebo Liu, Liang Ding, Miao Zhang, Jie Liu, Min Zhang

Multi-agent systems (MAS) based on large language models (LLMs) have demonstrated significant potential in collaborative problem-solving. However, they still face substantial challenges of low communication efficiency and suboptimal task performance, making the careful design of the agents' communication topologies particularly important. Inspired by the management theory that roles in an efficient team are often dynamically adjusted, we propose \textbf{AgentDropout}, which identifies redundant agents and communication across different communication rounds by optimizing the adjacency matrices of the communication graphs and eliminates them to enhance both token efficiency and task performance. Compared to state-of-the-art methods, AgentDropout achieves an average reduction of 21.6% in prompt token consumption and 18.4% in completion token consumption, along with a performance improvement of 1.14 on the tasks. Furthermore, the extended experiments demonstrate that AgentDropout achieves notable domain transferability and structure robustness, revealing its reliability and effectiveness. We release our code at https://github.com/wangzx1219/AgentDropout.

Dialogue-RAG: Enhancing Retrieval for LLMs via Node-Linking Utterance Rewriting

Qiwei Li, Teng Xiao, Zuchao Li, Ping Wang, Mengjia Shen, hai zhao

Large Language Models (LLMs) and Retrieval Augmented Generation (RAG) methods have demonstrated significant potential on tasks across multiple domains. However, ellipses and coreferences, as common phenomena in dialogue scenes, pose challenges to LLMs' understanding and RAG's retrieval accuracy. The previous works ignore the negative impact of this fuzzy data on RAG system. We explore the capabilities of LLMs and RAG systems in dialogue scenarios and use Incomplete Utterance Rewriting (IUR) to complete the key information in dialogue to enhance retrieval. Besides, we propose a lightweight IUR model for query rewriting. It is an end-to-end framework for node linking and iterative inference, incorporating two newly proposed probing semantic features derived from generative pre-training. This framework treats IUR as a series of link decisions on the input sequence and the incrementally constructed rewriting outputs. To test the performance of RAG system in the model multi-round dialogue scenario, we construct an RAG dialogue dataset on English and Chinese, Dialogue-RAG-MULTI-v1.0. Experiment results show that utterance rewriting can effectively improve the retrieval and generation ability of RAG system in dialogue scenes. Experiments on IUR tasks demonstrate the excellent performance of our lightweight IUR method.

SOTOPIA-: Dynamic Strategy Injection Learning and Social Instruction Following Evaluation for Social Agents

Wenyuan Zhang, Tianyun Liu, Mengxiao Song, Xiaodong Li, Tingwen Liu

Despite the abundance of prior social strategies possessed by humans, there remains a paucity of research dedicated to their transfer and integration into social agents. Our proposed SOTOPIA- framework aims to address and bridge this gap, with a particular focus on enhancing the social capabilities of language agents. This framework dynamically injects a variety of social strategies into expert agents, thereby automating the construction of high-quality social dialogue training corpus. Additionally, we introduce the concept of Social Instruction Following (S-IF) and propose two new S-IF evaluation metrics that are complementary to social capability. We demonstrate that several 7B models trained on high-quality corpus not only significantly surpasses the expert agent (GPT-4) in achieving social goals but also enhances S-IF performance. Analysis and variant experiments validate the advantages of dynamic construction, which can especially break the agent's prolonged deadlock.

OmniCharacter: Towards Immersive Role-Playing Agents with Seamless Speech-Language Personality Interaction

Haonan Zhang, Run Luo, Xiong Liu, Yuchuan Wu, Ting-En Lin, Pengpeng Zeng, QIANG QU, Feiteng Fang, Min Yang, Lianli Gao, Jingkuan Song, Fei Huang, Yongbin Li

Role-Playing Agents (RPAs), benefiting from large language models, is an emerging interactive AI system that simulates roles or characters with diverse personalities. However, existing methods primarily focus on mimicking dialogues among roles in textual form, neglecting the role's voice traits (e.g., voice style and emotions) as playing a crucial effect in interaction, which tends to be more immersive experiences in realistic scenarios. Towards this goal, we propose OmniCharacter, a first seamless speech-language personality interaction model to achieve immersive RPAs with low latency. Specifically, OmniCharacter enables agents to consistently exhibit role-specific personality traits and vocal traits throughout the interaction, enabling a mixture of speech and language responses. To align the model with speech-language scenarios, we construct a dataset named OmniCharacter-10K, which involves more distinctive characters (20), richly contextualized multi-round dialogue (10K), and dynamic speech response (135K). Experimental results showcase that our method yields better responses in terms of both content and style compared to existing RPAs and mainstream speech-language models, with a response latency as low as 289ms.

Controllable and Reliable Knowledge-Intensive Task Agents with Declarative GenieWorksheets

Harshit Joshi, Shicheng Liu, James Chen, Larsen Weigle, Monica Lam

Large Language Models are capable of carrying out human-like conversations in diverse settings in response to user requests for tasks and knowledge. However, existing conversational agents implemented with LLMs often struggle with hallucination, following instructions with conditional logic, and integrating knowledge from different sources. These shortcomings compromise the agents' effectiveness, rendering them unsuitable for deployment. To address these challenges, we introduce Genie, a programmable framework for creating knowledge-intensive task-oriented conversational agents that handle involved interactions and answer complex queries. Unlike LLMs, Genie delivers reliable, grounded responses through advanced dialogue state management and supports controllable agent policies via its declarative specification – Genie Worksheet. This is achieved through an algorithmic runtime system that implements the developer-supplied policy, limiting LLMs to (1) parse user input using a succinct conversational history, and (2) generate responses according to supplied content. Agents built with Genie outperform SOTA methods on complex logic dialogue datasets by up to 20.5%. We conducted a user study with 62 participants. Genie agents with GPT-4 Turbo outperformed the GPT-4 Turbo agents with function calling, improving goal completion rates from 21.8% to 82.8% across three real-world tasks.

Improving Dialogue State Tracking through Combinatorial Search for In-Context Examples

Haesung Pyun, Yoonah Park, Yohan Jo

In dialogue state tracking (DST), in-context learning comprises a retriever that selects labeled dialogues as in-context examples and a DST model that uses these examples to infer the dialogue state of the query dialogue. Existing methods for constructing training data for retrievers suffer from three key limitations: (1) the synergistic effect of examples is not considered, (2) the linguistic characteristics of the query are not sufficiently factored in, and (3) scoring is not directly optimized for DST performance. Consequently, the retriever can fail to retrieve examples that would substantially improve DST performance. To address these issues, we present CombiSearch—a method that scores effective in-context examples based on their combinatorial impact on DST performance. Our evaluation on MultiWOZ shows that retrievers trained with CombiSearch surpass state-of-the-art models, achieving a 20× gain in data efficiency and generalizing well to the SGD dataset. Moreover, CombiSearch attains a 12\% absolute improvement in the upper bound DST performance over traditional approaches when no retrieval errors are assumed. This significantly increases the headroom for practical DST performance while demonstrating that existing methods rely on suboptimal data for retriever training.

Dialogue Systems for Emotional Support via Value Reinforcement

Juhee Kim, Chunghu Mok, Jisun LEE, Hyang Sook Kim, Yohan Jo

Emotional support dialogue systems aim to reduce help-seekers' distress and help them overcome challenges. While human values—core beliefs that shape an individual's priorities—are increasingly emphasized in contemporary psychological therapy for their role in fostering internal transformation and long-term emotional well-being, their integration into emotional support systems remains underexplored. To bridge this gap, we present a value-driven method for training emotional support dialogue systems designed to reinforce positive values in seekers. Notably, our model identifies which values to reinforce at each turn and how to do so, by leveraging online support conversations from Reddit. We evaluate the method across support skills, seekers' emotional intensity, and value reinforcement. Our method consistently outperforms various baselines, effectively exploring and eliciting values from seekers. Additionally, leveraging crowd knowledge from Reddit significantly enhances its effectiveness. Therapists highlighted its ability to validate seekers' challenges and emphasize positive aspects of their situations—both crucial elements of value reinforcement. Our work, being the first to integrate value reinforcement into emotional support systems, demonstrates its promise and establishes a foundation for future research.

Divide-Then-Aggregate: An Efficient Tool Learning Method via Parallel Tool Invocation

Dongsheng Zhu, Weixian Shi, Zhengliang Shi, Zhaochun Ren, Shuaiqiang Wang, Lingyong Yan, Dawei Yin

While Large Language Models (LLMs) demonstrate remarkable capabilities, their ability to autonomously execute complex real-world tasks remains limited. Accordingly, tool learning has emerged to enable LLMs to effectively leverage external tools to extend their capabilities. Current tool-learning paradigms like CoT/ReAct employ sequential tool invocation but suffer from constrained perception and inadequate task planning. Alternative approaches using search-based decision trees incur substantial computational overhead. To address these limitations, we propose DTA-Llama (Divide-Then-Aggregate Llama), a novel parallel tool invocation framework featuring: (1) A Directed Acyclic Graph (DAG) structure that transformed from traditional tree-based tool search paths, enabling parallel execution and contributing high-quality training data; (2) A process-thread-inspired inference mechanism that iteratively decomposes tasks into parallel tool-using subtasks while aggregating results for subsequent decisions. Experimental results show that our approach substantially enhances task performance while reducing token consumption and inference time. Llama2-7B, using our method, is comparable to the official parallel function calling method of GPT-3.5. The relevant code, dataset, and model weights are available at https://corn0205.github.io/.

Enabling Chatbots with Eyes and Ears: An Immersive Multimodal Conversation System for Dynamic Interactions

Jihyoung Jang, Minwook Bae, Minji Kim, Dilek Hakkani-Tür, Hyounghun Kim

As chatbots continue to evolve toward human-like, real-world, interactions, multimodality remains an active area of research and exploration. So far, efforts to integrate multimodality into chatbots have primarily focused on image-centric tasks, such as visual dialogue and image-based instructions, placing emphasis on the "eyes" of human perception while neglecting the "ears", namely auditory aspects. Moreover, these studies often center around static interactions that focus on discussing the modality rather than naturally incorporating it into the conversation, which limits the richness of simultaneous, dynamic engagement. Furthermore, while multimodality has been explored in multi-party and multi-session conversations, task-specific constraints have hindered its seamless integration into dynamic, natural conversations. To address these challenges, this study aims to equip chatbots with "eyes and ears" capable of more immersive interactions with humans. As part of this effort, we introduce a new multimodal conversation dataset, Multimodal Multi-Session Multi-Party Conversation (M^3C), and propose a novel multimodal conversation model featuring multimodal memory retrieval. Our model, trained on the M^3C, demonstrates the ability to seamlessly engage in long-term conversations with multiple speakers in complex, real-world-like settings, effectively processing visual and auditory inputs to understand and respond appropriately. Human evaluations highlight the model's strong performance in maintaining coherent and dynamic interactions, demonstrating its potential for advanced multimodal conversational agents.

AmbiK: Dataset of Ambiguous Tasks in Kitchen Environment

Anastasia Ivanova, Zoya Volovikova, Bakaeva Eva, Alexey Kovalev, Aleksandr Panov

As a part of an embodied agent, Large Language Models (LLMs) are typically used for behavior planning given natural language instructions from the user. However, dealing with ambiguous instructions in real-world environments remains a challenge for LLMs. Various methods for task ambiguity detection have been proposed. However, it is difficult to compare them because they are tested on different datasets and there is no universal benchmark. For this reason, we propose AmbiK (Ambiguous Tasks in Kitchen Environment), the fully textual dataset of ambiguous instructions addressed to a robot in a kitchen environment. AmbiK was collected with the assistance of LLMs and is human-validated. It comprises 1000 pairs of ambiguous tasks and their unambiguous counterparts, categorized by ambiguity type (Human Preferences, Common Sense Knowledge, Safety), with environment descriptions, clarifying questions and answers, user intents, and task plans, for a total of 2000 tasks. We hope that AmbiK will enable researchers to perform a unified comparison of ambiguity detection methods. AmbiK is available at https://github.com/cog-model/AmbiK-dataset.

Session 2: Discourse and Pragmatics (Monday 28 July, 11:00-12:30, Hall 4/5)

Can LLMs Ground when they (Don't) Know: A Study on Direct and Loaded Political Questions

Clara Lachenmaier, Judith Sieker, Sina Zarrieß

Communication among humans relies on conversational grounding, allowing interlocutors to reach mutual understanding even when they do not have perfect knowledge and must resolve discrepancies in each other's beliefs. This paper investigates how large language models (LLMs) manage common ground in cases where they (don't) possess knowledge, focusing on facts in the political domain where the risk of misinformation and grounding failure is high. We examine LLMs' ability to answer direct knowledge questions and loaded questions that presuppose misinformation. We evaluate whether loaded questions lead LLMs to engage in active grounding and correct false user beliefs, in connection to their level of knowledge and their political bias. Our findings highlight significant challenges in LLMs' ability to engage in grounding and reject false user beliefs, raising concerns about their role in mitigating misinformation in political discourse.

Session 2: Efficient/Low-Resource Methods for NLP (Monday 28 July, 11:00-12:30, Hall 4/5)

TrimLLM: Progressive Layer Dropping for Domain-Specific LLMs

Lanxiang Hu, Tajana Rosing, Hao Zhang

Specializing large language models (LLMs) for local deployment in domain-specific use cases is necessary for strong performance while meeting latency and privacy constraints. However, conventional task-specific adaptation approaches do not show simultaneous memory saving and inference speedup at deployment time. Practical compression techniques like quantization and pruning require dedicated hardware or kernel support to achieve measured inference speedup. We develop TrimLLM based on the layer-wise specialization phenomenon we empirically observed and verified on contemporary LLMs. TrimLLM reduces the depth of LLMs via progressive layer dropping. We show it retains LLMs' capacity in specific domains and achieves inference speedup irrespective of hardware and deep learning frameworks. We evaluated TrimLLM on LLMs of various sizes for inference; models adapted on medical, legal, and financial datasets all demonstrate 2.1 - 5.7\times inference speedup on consumer GPUs and up to 3.1\times speedup on A100 when compared to state-of-the-art model compression algorithms, with no loss in accuracy at 50\sim 60\% model compression ratio.

KV-Latent: Dimensional-level KV Cache Reduction with Frequency-aware Rotary Positional Embedding

Shi Luohe, Zuchao Li, Lefei Zhang, Baoyuan Qi, Liu Guoming, hai zhao

Large language models (LLMs) based on Transformer Decoders have become the preferred choice for conversational generative AI. Despite the overall superiority of the Decoder architecture, the gradually increasing Key-Value (KV) cache during inference has emerged as a primary efficiency bottleneck, both in aspects of memory consumption and data transfer bandwidth limitations. To address these challenges, we propose a paradigm called KV-Latent. By down-sampling the Key-Value vector dimensions into a latent space, we can significantly reduce the KV Cache footprint and improve inference speed, only with a small amount of extra training, less than 1% of pre-training takes. Besides, we enhanced the stability of Rotary Positional Embedding applied on lower-dimensional vectors by modifying its frequency sampling mechanism, avoiding noise introduced by higher frequencies while retaining position attenuation. Our experiments, including both models with Grouped Query Attention and those without, have yielded satisfactory results. Finally, we conducted comparative experiments to study the impact of separately reducing Key and Value components on model's performance. Our approach allows for the construction of more efficient language model systems, and opens the new possibility on KV Cache saving and efficient LLMs.

WarriorCoder: Learning from Expert Battles to Augment Code Large Language Models

Huawen Feng, Pu Zhao, Qingfeng Sun, Can Xu, Fangkai Yang, Lu Wang, Qianli Ma, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, Qi Zhang

Despite recent progress achieved by code large language models (LLMs), their remarkable abilities are largely dependent on fine-tuning on the high-quality data, posing challenges for data collection and annotation. To address this, current methods often design various data flywheels to collect complex code instructions, enabling models to handle more intricate tasks. However, these approaches typically rely on off-the-shelf datasets and data augmentation from a limited set of proprietary LLMs (e.g., Claude, GPT4, and so on), which restricts the diversity of the constructed data and makes it prone to systemic biases. In this paper, we propose **WarriorCoder**, a novel paradigm learns from expert battles to address these limitations. Specifically, we create an arena where leading expert code LLMs challenge each other, with evaluations conducted by impartial judges. This competitive framework generates novel training data from scratch, leveraging the strengths of all participants. Experimental results show that **WarriorCoder** achieves state-of-the-art performance compared to previous models of the same size, even without relying on proprietary LLMs.

P^2 Law: Scaling Law for Post-Training After Model Pruning

Xiaodong Chen, Yuxuan Hu, Xiaokang Zhang, Yanling Wang, Cuiping Li, Hong Chen, Jing Zhang

Pruning has become a widely adopted technique for reducing the hardware requirements of large language models (LLMs). To recover model performance after pruning, post-training is commonly employed to mitigate the resulting performance degradation. While post-training benefits from larger datasets, once the dataset size is already substantial, increasing the training data provides only limited performance gains. To balance post-training cost and model performance, it is necessary to explore the optimal amount of post-training data. Through extensive experiments on the Llama-3 and Qwen-2.5 series models, pruned using various common pruning methods, we uncover the scaling \textbf{Law} for \textbf{P}ost-training after model \textbf{P}runing, referred to as the P^2 Law. This law identifies four key factors for predicting the pruned model's post-training loss: the model size before pruning, the number of post-training tokens, the pruning rate, and the model's loss before pruning. Moreover, P^2 Law can generalize to larger dataset sizes, larger model sizes, and higher pruning rates, offering valuable insights for the post-training of pruned LLMs.

CoT-Valve: Length-Compressible Chain-of-Thought Tuning

Xinyin Ma, Guangnian Wan, Runpeng Yu, Gongfan Fang, Xinchao Wang

Chain-of-Thought significantly enhances a model's reasoning capability, but it also comes with a considerable increase in inference costs due to long chains. With the observation that the reasoning path can be easily compressed under easy tasks but struggle on hard tasks, we explore the feasibility of elastically controlling the length of reasoning paths with only one model, thereby reducing the inference overhead of reasoning models dynamically based on task difficulty. We introduce a new tuning and inference strategy named CoT-Valve, designed to allow models to generate reasoning chains of varying lengths. To achieve this, we propose to identify a direction in the parameter space that, when manipulated, can effectively control the length of generated CoT. Moreover, we show that this property is valuable for compressing the reasoning chain. We construct datasets with chains from long to short for the same questions and explore two enhanced strategies for CoT-Valve: (1) a precise length-compressible CoT tuning method, and (2) a progressive chain length compression approach. Our experiments show that CoT-Valve successfully enables controllability and compressibility of the chain and shows better performance than the prompt-based control. We applied this method to QwQ-32B-Preview, reducing reasoning chains on GSM8K from 741 to 225 tokens with a minor performance drop (95.07% to 94.92%) and on AIME from 6827 to 4629 tokens, with only one additional incorrect answer.

Combining Domain and Alignment Vectors Provides Better Knowledge-Safety Trade-offs in LLMs

Megh Thakkar, Quentin Fournier, Matthew Riemer, Pin-Yu Chen, Amal Zouaq, Payel Das, Sarath Chandar

There is a growing interest in training domain-expert LLMs that excel in specific technical fields compared to their general-purpose instruction-tuned counterparts. However, these expert models are not either explicitly trained to be safe, or experience a loss in their safety abilities in the process, making them capable of generating harmful content. We observe that simple interpolation between the domain and alignment delta parameters leads to safer domain-specific models that preserve their utility. Building on this, we introduce MergeAlign, a simple, efficient, and effective model merging-based alignment method. We apply MergeAlign on Llama3 models that are experts in medicine and finance, obtaining substantial safety alignment improvements with minimal to no degradation on domain-specific benchmarks. We study the impact of model merging through model similarity metrics and contributions of individual models being merged, as well as the applicability of MergeAlign on more general code and math expert models using the Qwen-2.5 series of models. We hope our findings open new research avenues towards efficient development and deployment of safe expert LLMs.

A Little Human Data Goes A Long Way

Dhananjay Ashok, Jonathan May

Faced with an expensive human annotation process, creators of NLP systems increasingly turn to synthetic data generation. While this method shows promise, the extent to which synthetic data can replace human annotation is poorly understood. We investigate the use of synthetic data in Fact Verification (FV) and Evidence-based Question Answering (QA) by incrementally replacing human-generated data with synthetic points on eight diverse datasets. Strikingly, replacing up to 90% of the training data only marginally decreases performance, but replacing the final 10% leads to severe declines. We find that models trained on purely synthetic data can be improved by including as few as 125 human-generated data points. We show that matching the performance gain of a little human data requires an order of magnitude more synthetic data, and then estimate price ratios at which human annotation would be a more cost-effective solution. Our results suggest that even when human annotation at scale is infeasible, there is great value to having a small proportion of the dataset being human-generated.

DavIR: Data Selection via Implicit Reward for Large Language Models

Haotian Zhou, Tingkai Liu, Qianli Ma, Yufeng Zhang, Jianbo Yuan, Pengfei Liu, Yang You, Hongxia Yang

We introduce DavIR, a model-based data selection method for post-training Large Language Models. DavIR generalizes Reducible Holdout Loss to core-set selection problem of causal language modeling, and quantifies the learnability of a given datum with respect to a pre-trained LLM based on relative reduction in loss during fine-tuning, a metric we show to be closely related to the implicit reward model described in Direct Preference Optimization (DPO). We show that 6% of Alpaca dataset selected with DavIR can steer both the LLaMA and Gemma model family to produce superior performance compared to the same models trained on the full 52K dataset. We also show that Alpaca dataset compressed with DavIR can be combined with GSM8K dataset to effectively balance open-domain freeform QA and mathematical reasoning capabilities. Finally, we apply the DavIR objective to DPO and develop a normalized DavIR-DPO objective which improves alignment performance of Zephyr-7B-SFT model by 8% (relative) on AlpacaEval, compared against training on vanilla DPO objective.

Forward Knows Efficient Backward Path: Saliency-Guided Memory-Efficient Fine-tuning of Large Language Models

Yeachan Kim, SangKeun Lee

Fine-tuning is widely recognized as a crucial process for aligning large language models (LLMs) with human intentions. However, the substantial memory requirements associated with fine-tuning pose a significant barrier to extending the applicability of LLMs. While parameter-efficient fine-tuning can be a promising approach by reducing trainable parameters, intermediate activations still need to be cached to compute gradients during the backward pass, thereby limiting overall memory efficiency. In this work, we propose Saliency-Guided Gradient Flow (SAGE), a memory-efficient fine-tuning method designed to minimize the memory specifically associated with cached intermediate activations. The key strategy is to selectively cache activations based on their saliency during the forward pass and then use these activations for the backward pass. This process transforms the dense backward pass into a sparse one, thereby enhancing memory efficiency. To verify whether SAGE can serve as an efficient alternative for fine-tuning, we conduct comprehensive experiments across diverse fine-tuning scenarios and setups. The experimental results show that SAGE substantially improves memory efficiency without a significant loss in accuracy, highlighting its broad value in real-world applications

LongRecipe: Recipe for Efficient Long Context Generalization in Large Language Models

Zhiyuan Hu, Yuliang Liu, Jinman Zhao, Suyuchen Wang, WangYan, Wei Shen, Qing Gu, Anh Tuan Luu, See-Kiong Ng, Zhiwei Jiang, Bryan Hooi

Large language models (LLMs) face significant challenges in handling long-context tasks because of their limited effective context window size during pretraining, which restricts their ability to generalize over extended sequences. Meanwhile, extending the context window in LLMs through post-pretraining is highly resource-intensive. To address this, we introduce LongRecipe, an efficient training strategy for extending the context window of LLMs, including impactful token analysis, position index transformation, and training optimization strategies. It simulates long-sequence inputs while maintaining training efficiency and significantly improves the model's understanding of long-range dependencies. Experiments on three types of LLMs show that LongRecipe can utilize long sequences while requiring only 30\% of the target context window size, and reduces computational training resource over 85\% compared to full sequence training. Furthermore, LongRecipe also preserves the original LLM's capabilities in general tasks. Ultimately, \textit{we can extend effective context window of open-source LLMs from 8k to 128k, achieving performance close to GPT-4 with just one day of dedicated training using a single GPU with 80G memory.} Our code is released at https://github.com/zhiyuanhubj/LongRecipe.

Outlier-Safe Pre-Training for Robust 4-Bit Quantization of Large Language Models

Jungwoo Park, Chanwoong Yoon, Hyeon Hwang, Taewhoo Lee, Jaewoo Kang

Extreme activation outliers in Large Language Models (LLMs) critically degrade quantization performance, hindering efficient on-device deployment. While channel-wise operations and adaptive gradient scaling are recognized causes, practical mitigation remains challenging. We introduce **Outlier-Safe Pre-Training (OSP)**, a practical guideline that proactively prevents outlier formation, rather than relying on post-hoc mitigation. OSP combines three key innovations: (1) the Muon optimizer, eliminating privileged bases while maintaining training efficiency, (2) Single-Scale RMSNorm, preventing channel-wise amplification, and (3) a learnable embedding projection, redistributing activation magnitudes. We validate OSP by training a 1.4B-parameter model on 1 trillion tokens, which is the first production-scale LLM trained without such outliers. Under aggressive 4-bit quantization, our OSP model achieves a 35.7 average score across 10 benchmarks (versus 26.5 for an Adam-trained model), with only a 2% training overhead. Remarkably, OSP models exhibit near-zero excess kurtosis (0.04) compared to extreme values (1818.56) in standard models, fundamentally altering LLM quantization behavior. Our work demonstrates that outliers are not inherent to LLMs but are consequences of training strategies, paving the way for more efficient LLM deployment.

State-offset Tuning: State-based Parameter-Efficient Fine-Tuning for State Space Models

Wonjun Kang, Kevin Galim, Yuchen Zeng, Minjae Lee, Hyung Il Koo, Nam Ik Cho

State Space Models (SSMs) have emerged as efficient alternatives to Transformers, mitigating their quadratic computational cost. However, the application of Parameter-Efficient Fine-Tuning (PEFT) methods to SSMs remains largely unexplored. In particular, prompt-based methods like Prompt Tuning and Prefix-Tuning, which are widely used in Transformers, do not perform well on SSMs. To address this, we propose **state-based methods** as a superior alternative to prompt-based methods. This new family of methods naturally stems from the architectural characteristics of SSMs. State-based methods adjust state-related features directly instead of depending on external prompts. Furthermore, we introduce a novel state-based PEFT method: **State-offset Tuning**. At every timestep, our method directly affects the state at the current step, leading to more effective adaptation. Through extensive experiments across diverse datasets, we demonstrate the effectiveness of our method. Code is available at https://github.com/furiosa-ai/ssm-state-tuning.

SCAR: Data Selection via Style Consistency-Aware Response Ranking for Efficient Instruction-Tuning of Large Language Models

Zhuang Li, YUNCHENG HUA, Thuy-Trang Vu, Haolan Zhan, Lizhen Qu, Gholamreza Haffari
Recent studies emphasize that manually ensuring a consistent response style and maintaining high data quality in training sets can significantly improve the performance of fine-tuned Large Language Models (LLMs) while reducing the number of training examples needed. However, the precise definition of style and the relationship between style, data quality, and LLM performance remains unclear. This research identifies two key stylistic elements in responses: linguistic form and instructional surprisal. We find that, among training data of comparable quality, higher consistency in these response elements leads to better LLM performance. Inspired by this, we introduce Style Consistency-Aware Response Ranking (SCAR), which automatically prioritizes instruction-response pairs in the training set based on their response stylistic consistency. By selecting the most style-consistent examples, using 0.7% of the full dataset in certain cases, the fine-tuned LLMs can match or even surpass the performance of models trained on the entire dataset in coding and open-ended question-answering benchmarks. Code and data are available at https://github.com/zhuang-li/SCAR .

Accurate KV Cache Quantization with Outlier Tokens Tracing

Yi Su, Yuechi Zhou, Quantong Qiu, Juntao Li, Qingrong Xia, Ping Li, Xinyu Duan, Zhefeng Wang, Min Zhang
The impressive capabilities of Large Language Models (LLMs) come at the cost of substantial computational resources during deployment. While KV Cache can significantly reduce recomputation during inference, it also introduces additional memory overhead. KV Cache quantization presents a promising solution, striking a good balance between memory usage and accuracy. Previous research has shown that the Keys are distributed by channel, while the Values are distributed by token. Consequently, the common practice is to apply channel-wise quantization to the Keys and token-wise quantization to the Values. However, our further investigation reveals that a small subset of unusual tokens exhibit unique characteristics that deviate from this pattern, which can substantially impact quantization accuracy. To address this, we develop a simple yet effective method to identify these tokens accurately during the decoding process and exclude them from quantization as outlier tokens, significantly improving overall accuracy. Extensive experiments show that our method achieves significant accuracy improvements under 2-bit quantization and can deliver a 6.4 times reduction in memory usage and a 2.3 times increase in throughput.

TeamLoRA: Boosting Low-Rank Adaptation with Expert Collaboration and Competition

Tianwei Lin, Jiang Liu, Wenqiao Zhang, Yang Dai, Haoyuan Li, Zhelun Yu, Wanggui He, Juncheng Li, Jiannan Guo, Hao Jiang, Siliang Tang, Yueting Zhuang
While Parameter-Efficient Fine-Tuning (PEFT) methods like Low-Rank Adaptation (LoRA) effectively address resource constraints during fine-tuning, their performance often falls short, especially in multidimensional task scenarios. To address this issue, one straightforward solution is to introduce task-specific LoRA as domain experts, leveraging the modeling of multiple capabilities of experts and thus enhancing the general capability of multi-task learning. Although promising, these additional components often add complexity to the training and inference process, contravening the efficiency that PEFT is designed to deliver. Considering this, we introduce an innovative PEFT method, **TeamLoRA**, consisting of a collaboration and competition module for LoRA experts, thus achieving the right balance of effectiveness and efficiency: **(i)** For *collaboration*, we introduce a novel knowledge sharing and organization mechanism designed to optimize hierarchical learning while enhancing the efficiency of model training and inference. **(ii)** For *competition*, we propose leveraging a game-theoretic interaction mechanism for experts, encouraging experts to transfer their domain-specific knowledge while facing diverse downstream tasks, thus enhancing the performance. By doing so, TeamLoRA elegantly connects the experts as a ``*Team*'' with internal collaboration and competition, enabling a faster and more accurate PEFT paradigm. Meanwhile, we curate a **Comprehensive Multi-Task Evaluation (CME)** benchmark to thoroughly assess the capability of multi-task learning. Experiments conducted on our CME and other benchmarks indicate the effectiveness and efficiency of TeamLoRA. Our project is available at https://github.com/DCDmllm/TeamLoRA.

STUN: Structured-Then-Unstructured Pruning for Scalable MoE Pruning

Jaeseong Lee, seung-won hwang, Aurick Qiao, Daniel F Campos, Zhewei Yao, Yuxiong He

Mixture-of-experts (MoEs) have been adopted for reducing inference costs by sparsely activating experts in large language models (LLMs). Despite these reductions, the massive number of parameters in MoEs still makes them expensive to serve. Conventionally, unstructured or structured pruning has been considered to reduce number of parameters. Our key contribution is exploring the interpolation between structured and unstructured pruning, to propose a novel structured-then-unstructured (STUN) approach outperforming both of structured or unstructured pruning, especially for MoEs. In the first stage, we show a scalable expert pruning with O(1) forward pass, unlike existing work requiring O(\frac{k^n}{\sqrt{n}}) forward passes for n experts that cannot scale for recent MoEs with hundreds of experts. We then show our expert-pruned MoEs are robust to unstructured pruning to follow. Experiments on Snowflake Arctic and Mixtral shows that our proposal is highly effective-- For Snowflake Arctic, a 480B-sized MoE with 128 experts, our method needs only one H100 and two hours to achieve nearly no loss in performance with 40% sparsity, even in generative tasks such as GSM8K, where state-of-the-art structured or unstructured pruning methods fail. The code is publicly available.

YESciEval: Robust LLM-as-a-Judge for Scientific Question Answering

Jennifer D'Souza, Hamed Babaei Giglou, Quentin Münch

Large Language Models (LLMs) drive scientific question-answering on modern search engines, yet their evaluation robustness remains underexplored. We introduce YESciEval, an open-source framework that combines fine-grained rubric-based assessment with reinforcement learning to mitigate optimism bias in LLM evaluators. We release multidisciplinary scienceQ\&A datasets, including adversarial variants, with evaluation scores from multiple LLMs. Independent of proprietary models and human feedback, our approach enables scalable, cost-free evaluation. By advancing reliable LLM-as-a-judge models, this work supports AI alignment and fosters robust, transparent evaluation essential for scientific inquiry.

MAmmoTH-VL: Eliciting Multimodal Reasoning with Instruction Tuning at Scale

Jiawei Guo, Tianyu Zheng, Yizhi LI, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Graham Neubig, Wenhu Chen, Xiang Yue

Open-source multimodal large language models (MLLMs) have shown significant potential in a broad range of tasks. However, their reasoning capabilities remain constrained by existing instruction-tuning datasets, which were predominately repurposed from academic datasets such as VQA, AI2D, and ChartQA. These datasets target simplistic tasks, and only provide phrase-level answers without any intermediate rationales.To address these challenges, we introduce a scalable and cost-effective method to construct a large-scale multimodal instruction-tuning dataset with rich intermediate rationales designed to elicit CoT reasoning. Using only open models, we create a dataset containing 12M instruction-response pairs to cover diverse reasoning-intensive tasks.Experiments demonstrate that training MLLMs on our dataset not only significantly improves reasoning capabilities, achieving state-of-the-art performance on benchmarks such as MathVerse (+8.1%), MMMU-Pro (+7%), and MuirBench (+13.3%), but also gains improvements of up to 4% on non-reasoning-based benchmarks.

Enhancing Cross-Lingual Transfer through Reversible Transliteration: A Huffman-Based Approach for Low-Resource Languages

Wenhao Zhuang, Yuan Sun, Xiaobing Zhao

As large language models (LLMs) are trained on increasingly diverse and extensive multilingual corpora, they demonstrate cross-lingual transfer capabilities. However, these capabilities often fail to effectively extend to low-resource languages, particularly those utilizing non-Latin scripts. While transliterating low-resource languages into Latin script presents a natural solution, there currently lacks a comprehensive framework for integrating transliteration into LLMs training and deployment. Taking a pragmatic approach, this paper innovatively combines character transliteration with Huffman coding to design a complete transliteration framework. Our proposed framework offers the following advantages: 1) Compression: Reduces storage requirements for low-resource language content, achieving up to 50% reduction in file size and 50-80% reduction in token count. 2) Accuracy: Guarantees 100% lossless conversion from transliterated text back to the source language. 3) Efficiency: Eliminates the need for vocabulary expansion for low-resource languages, improving training and inference efficiency. 4) Scalability: The framework can be extended to other low-resource languages. We validate the effectiveness of our framework across multiple downstream tasks, including text classification, machine reading comprehension, and machine translation. Experimental results demonstrate that our method significantly enhances the model's capability to process low-resource languages while maintaining performance on high-resource languages. Our data and code are publicly available at https://github.com/CMLI-NLP/HuffmanTranslit.

PROPER: A Progressive Learning Framework for Personalized Large Language Models with Group-Level Adaptation

Linhai Zhang, Jialong Wu, Deyu Zhou, Yulan He

Personalized large language models (LLMs) aim to tailor their outputs to user preferences. Recent advances in parameter-efficient fine-tuning (PEFT) methods have highlighted the effectiveness of adapting population-level LLMs to personalized LLMs by fine-tuning user-specific parameters with user history. However, user data is typically sparse, making it challenging to adapt LLMs to specific user patterns. To address this challenge, we propose PROgressive PERsonalization (PROPER), a novel progressive learning framework inspired by meso-level theory in social science. PROPER bridges population-level and user-level models by grouping users based on preferences and adapting LLMs in stages. It combines a Mixture-of-Experts (MoE) structure with Low Ranked Adaptation (LoRA), using a user-aware router to assign users to appropriate groups automatically. Additionally, a LoRA-aware router is proposed to facilitate the integration of individual user LoRAs with the group-level LoRA. Experimental results show that PROPER significantly outperforms SOTA models across multiple tasks, demonstrating the effectiveness of our approach.

Aligning Large Language Models to Follow Instructions and Hallucinate Less via Effective Data Filtering

Shuzheng Si, Haozhe Zhao, Gang Chen, Cheng Gao, Yuzhuo Bai, Zhitong Wang, Kaikai An, Kangyang Luo, Chen Qian, Fanchao Qi, Baobao Chang, Maosong Sun

Training LLMs on data containing unfamiliar knowledge during the instruction tuning stage can encourage hallucinations. To address this challenge, we introduce NOVA, a novel framework designed to identify high-quality data that aligns well with the LLM's learned knowledge to reduce hallucinations. NOVA includes Internal Consistency Probing (ICP) and Semantic Equivalence Identification (SEI) to measure how familiar the LLM is with instruction data. Specifically, ICP evaluates the LLM's understanding of the given instruction by calculating the tailored consistency among multiple self-generated responses. SEI further assesses the familiarity of the LLM with the target response by comparing it to the generated responses, using the proposed semantic clustering and well-designed voting strategy. Finally, to ensure the quality of selected samples, we introduce an expert-aligned reward model, considering characteristics beyond just familiarity. By considering data quality and avoiding unfamiliar data, we can utilize the selected data to effectively align LLMs to follow instructions and hallucinate less. Experiments show that NOVA significantly reduces hallucinations while maintaining a competitive ability to follow instructions.

Decoding by Contrasting Knowledge: Enhancing Large Language Model Confidence on Edited Facts

Baolong Bi, Shenghua Liu, Lingrui Mei, Yiwei Wang, Junfeng Fang, Pengliang Ji, Xueqi Cheng

The knowledge within large language models (LLMs) may become outdated quickly. While in-context editing (ICE) is currently the most effective method for knowledge editing (KE), it is constrained by the black-box modeling of LLMs and thus lacks interpretability. Our work aims to elucidate the superior performance of ICE in KE by analyzing the impacts of in-context new knowledge on token-wise distributions. We observe that despite a significant boost in logits of the new knowledge, the performance of ICE is still hindered by stubborn knowledge. We propose a novel approach termed Decoding by Contrasting Knowledge (DeCK). DeCK derives the distribution of the next token by contrasting the logits obtained from the newly edited knowledge guided by ICE with those from the unedited parametric knowledge. Our experiments demonstrate that DeCK enhances the confidence of LLMs in edited facts. For instance, it improves the performance of LLaMA3-8B-instruct on MQuAKE by up to 219%, demonstrating its capability to strengthen ICE. DeCK can be easily integrated into any ICE method as a decoding component to enhance editing capabilities.

Measuring Data Diversity for Instruction Tuning: A Systematic Analysis and A Reliable Metric

Yuming Yang, Yang Nan, Junjie Ye, Shihan Dou, Xiao Wang, Shuo Li, Huijie Lv, Tao Gui, Qi Zhang, Xuanjing Huang

Data diversity is crucial for the instruction tuning of large language models. Existing studies have explored various diversity-aware data selection methods to construct high-quality datasets and enhance model performance. However, the fundamental problem of precisely defining and measuring data diversity remains underexplored, limiting clear guidance for data engineering. To address this, we systematically analyze 11 existing diversity measurement methods by evaluating their correlation with model performance through extensive fine-tuning experiments. Our results indicate that a reliable diversity measure should properly account for both inter-sample differences and the information density in the sample space. Building on this, we propose NovelSum, a new diversity metric based on sample-level "novelty." Experiments on both simulated and real-world data show that NovelSum accurately captures diversity variations and achieves a 0.97 correlation with instruction-tuned model performance, highlighting its value in guiding data engineering practices. With NovelSum as an optimization objective, we further develop a greedy, diversity-oriented data selection strategy that outperforms existing approaches, validating both the effectiveness and practical significance of our metric.

ImPart: Importance-Aware Delta-Sparsification for Improved Model Compression and Merging in LLMs

Yan Yang, Yixia Li, Hongru WANG, Xuetao Wei, James Jianqiao Yu, Yun Chen, Guanhua Chen

With the proliferation of task-specific large language models, delta compression has emerged as a method to mitigate the resource challenges of deploying numerous such models by effectively compressing the delta model parameters. Previous delta-sparsification methods either remove parameters randomly or truncate singular vectors directly after singular value decomposition (SVD). However, these methods either disregard parameter importance entirely or evaluate it with too coarse a granularity. In this work, we introduce ImPart, a novel importance-aware delta sparsification approach. Leveraging SVD, it dynamically adjusts sparsity ratios of different singular vectors based on their importance, effectively retaining crucial task-specific knowledge even at high sparsity ratios. Experiments show that ImPart achieves state-of-the-art delta sparsification performance, demonstrating 2\times higher compression ratio than baselines at the same performance level. When integrated with existing methods, ImPart sets a new state-of-the-art on delta quantization and model merging.

Maximizing the Effectiveness of Larger BERT Models for Compression

Wen-Shu Fan, Su Lu, Shangyu Xing, Xin-Chun Li, De-Chuan Zhan

Knowledge distillation (KD) is a widely used approach for BERT compression, where a larger BERT model serves as a teacher to transfer knowledge to a smaller student model. Prior works have found that distilling a larger BERT with superior performance may degrade student's performance than a smaller BERT. In this paper, we investigate the limitations of existing KD methods for larger BERT models. Through Canonical Correlation Analysis, we identify that these methods fail to fully exploit the potential advantages of larger teachers. To address this, we propose an improved distillation approach that effectively enhances knowledge transfer. Comprehensive experiments demonstrate the effectiveness of our method in enabling larger BERT models to distill knowledge more efficiently.

Look Both Ways and No Sink: Converting LLMs into Text Encoders without Training

Ziyong Lin, Haoyi Wu, Shu Wang, Kewei Tu, Zilong Zheng, Zixia Jia

Recent advancements have demonstrated the advantage of converting pretrained large language models into powerful text encoders by enabling bidirectional attention in transformer layers. However, existing methods often require extensive training on large-scale datasets, posing challenges in low-resource, domain-specific scenarios. In this work, we show that a pretrained large language model can be converted into a strong text encoder without additional training. We first conduct a comprehensive empirical study to investigate different conversion strategies and identify the impact of the attention sink phenomenon on the performance of converted encoder models. Based on our findings, we propose a novel approach that enables bidirectional attention and suppresses the attention sink phenomenon, resulting in superior performance. Extensive experiments on multiple domains demonstrate the effectiveness of our approach. Our work provides new insights into the training-free conversion of text encoders in low-resource scenarios and contributes to the advancement of domain-specific text representation generation. Our code is available at https://github.com/bigai-nlco/Look-Both-Ways-and-No-Sink.

Data Whisperer: Efficient Data Selection for Task-Specific LLM Fine-Tuning via Few-Shot In-Context Learning

Shaobo Wang, Xiangqi Jin, Ziming Wang, Jize Wang, Jiajun Zhang, Kaixin Li, Zichen Wen, Zhong Li, Conghui He, Xuming Hu, Linfeng Zhang

Fine-tuning large language models (LLMs) on task-specific data is essential for their effective deployment. As dataset sizes grow, efficiently selecting optimal subsets for training becomes crucial to balancing performance and computational costs. Traditional data selection methods often require fine-tuning a scoring model on the target dataset, which is time-consuming and resource-intensive, or rely on heuristics that fail to fully leverage the model's predictive capabilities. To address these challenges, we propose Data Whisperer, an efficient, training-free, attention-based method that leverages few-shot in-context learning with the model to be fine-tuned. Comprehensive evaluations were conducted on both raw and synthetic datasets across diverse tasks and models. Notably, Data Whisperer achieves superior performance compared to the full GSM8K dataset on the Llama-3-8B-Instruct model, using just 10% of the data, and outperforms existing methods with a 3.1-point improvement and a 7.4× speedup.

MobiLoRA: Accelerating LoRA-based LLM Inference on Mobile Devices via Context-aware KV Cache Optimization

Borui Li, Yitao Wang, Haoran Ma, Ligeng Chen, Jun Xiao, Shuai Wang

Deploying large language models (LLMs) with low-rank adaptation (LoRA) on mobile devices is promising due to their capability to complete diverse domain-specific tasks while ensuring privacy and accessibility. In this paper, we introduce MobiLoRA to accelerate LoRA-based LLM inference on mobile devices. MobiLoRA focuses on optimizing the key-value (KV) caches due to the limited computing and memory resources of mobile devices. The key insight of MobiLoRA lies in the utilization of two contexts for on-device LoRA serving: semantic-level contexts, such as prompts with shared prefixes, and system-level contexts, such as the application status (e.g., foreground or killed) of LLM requests. Specifically, for semantic-level contexts, MobiLoRA proposes similarity-aware delta encoding, which leverages token-wise similarity in KV caches across LoRA adapters for efficient storage and reuse. Furthermore, MobiLoRA advocates context-aware KV cache management to optimize cache retention and eviction considering the system-level contexts. We fully implement MobiLoRA and compare it with state-of-the-art LLM serving frameworks using real-world mobile device traces. Results show that MobiLoRA accelerates LoRA-based LLM inference by 57.6% on mobile devices.

RefreshKV: Updating Small KV Cache During Long-form Generation

Fangyuan Xu, Tanya Goyal, Eunsol Choi

Generating long sequences of tokens given a long-context input is a very compute-intensive inference scenario for large language models (LLMs). One prominent inference speed-up approach is constructing a smaller key-value (KV) cache, relieving LLMs from computing attention over a long sequence of tokens. While such methods work well to generate short sequences, their performance degrades rapidly for long-form generation. Most KV compression happens once, prematurely removing tokens that can be useful later in the generation. We propose a new inference-time method, RefreshKV, that flexibly alternates between full context attention and attention over a subset of input tokens during generation. After each full attention step, we update the smaller KV cache based on the attention pattern over the entire input. Applying our method to off-the-shelf LLMs achieves comparable speedup to eviction-based methods while improving performance for various long-form generation tasks. Lastly, we show that continued pretraining with our inference setting brings further gains in performance.

Mixtures of In-Context Learners

Giwon Hong, Emile van Krieken, Edoardo Ponti, Nikolay Malkin, Pasquale Minervini

In-context learning (ICL) adapts LLMs by providing demonstrations without fine-tuning the model parameters; however, it is very sensitive to the choice of in-context demonstrations, and processing many demonstrations can be computationally demanding. We propose Mixtures of In-Context Learners (MoICL), a novel approach that uses subsets of demonstrations to train a set of experts via ICL and learns a weighting function to merge their output distributions via gradient-based optimisation. In our experiments, we show performance improvements on 5 out of 7 classification datasets compared to a set of strong baselines (e.g., up to +13% compared to ICL and LENS). Moreover, we improve the Pareto frontier of ICL by reducing the inference time needed to achieve the same performance with fewer demonstrations. Finally, MoICL is more robust to out-of-domain (up to +11%), imbalanced (up to +49%) and perturbed demonstrations (up to +38%).

SpindleKV: A Novel KV Cache Reduction Method Balancing Both Shallow and Deep Layers

Zicong Tang, Shi Luohe, Zuchao Li, Baoyuan Qi, Liu Guoming, Lefei Zhang, Ping Wang

Large Language Models (LLMs) have achieved impressive accomplishments in recent years. However, the increasing memory consumption of KV cache has possessed a significant challenge to the inference system. Eviction methods have revealed the inherent redundancy within the KV cache, demonstrating its potential for reduction, particularly in deeper layers. However, KV cache reduction for shallower layers has been found to be insufficient. Based on our observation that, the KV cache exhibits a high degree of similarity. Based on this observation, we proposed a novel KV cache reduction method, SpindleKV, which balances both shallow and deep layers. For deep layers, we employ an attention weight based eviction method, while for shallow layers, we apply a codebook based replacement approach which is learnt by similarity and merging policy. Moreover, SpindleKV addressed the Grouped-Query Attention (GQA) dilemma faced by other attention based eviction methods. Experiments on two common benchmarks with three different LLMs shown that SpindleKV obtained better KV cache reduction effect compared to baseline methods, while preserving similar or even better model performance.

Unifying Uniform and Binary-coding Quantization for Accurate Compression of Large Language Models

Seungcheol Park, Jeongin Bae, Beomseok Kwon, Minjun Kim, Byeongwook Kim, Se Jung Kwon, U Kang, Dongsoo Lee

How can we quantize large language models while preserving accuracy? Quantization is essential for deploying large language models (LLMs) efficiently. Binary-coding quantization (BCQ) and uniform quantization (UQ) are promising quantization schemes that have strong expressiveness and optimizability, respectively. However, neither scheme leverages both advantages. In this paper, we propose UniQuan_F (Unified Quantization with Flexible Mapping), an accurate quantization method for LLMs. UniQuan_F harnesses both strong expressiveness and optimizability by unifying the flexible mapping technique in UQ and BCQ’s non-uniform quantization levels. We propose unified initialization, and local and periodic mapping techniques to optimize the parameters in UniQuan_F precisely. After optimization, our unification theorem removes computational and memory overhead, allowing us to utilize the superior accuracy of UniQuan_F without extra deployment costs induced by the unification. Experimental results demonstrate that UniQuan_F outperforms existing UQ and BCQ methods, achieving up to 4.60% higher accuracy on GSM8K benchmark.

Multi-Facet Blending for Faceted Query-by-Example Retrieval

Heejin Do, Sangwon Ryu, Jonghwi Kim, Gary Lee

With the growing demand to fit fine-grained user intents, faceted query-by-example (QBE), which retrieves similar documents conditioned on specific facets, has gained recent attention. However, prior approaches mainly depend on document-level comparisons using basic indicators like citations due to the lack of facet-level relevance datasets; yet, this limits their use to citation-based domains and fails to capture the intricacies of facet constraints. In this paper, we propose a multi-facet blending (FaBle) augmentation method, which exploits modularity by decomposing and recomposing to explicitly synthesize facet-specific training sets. We automatically decompose documents into facet units and generate (ir)relevant pairs by leveraging LLMs' intrinsic distinguishing capabilities; then, dynamically recomposing the units leads to facet-wise relevance-informed document pairs. Our modularization eliminates the need for pre-defined facet knowledge or labels. Further, to prove the FaBle's efficacy in a new domain beyond citation-based scientific paper retrieval, we release a benchmark dataset for educational exam item QBE. FaBle augmentation on 1K documents remarkably assists training in obtaining facet conditional embeddings.

Pretraining Context Compressor for Large Language Models with Embedding-Based Memory

Yuhong Dai, Jianxun Lian, Yitian Huang, Wei Zhang, Mingyang Zhou, Mingqi Wu, Xing Xie, Hao Liao

Efficient processing of long contexts in large language models (LLMs) is essential for real-world applications like retrieval-augmented generation and in-context learning, especially in resource-constrained environments such as edge computing. This paper explores the embedding-based context compression to reduce inference costs while preserving the downstream LLM configurations. We propose a decoupled compressor-LLM framework, pretrained on text reconstruction and completion tasks, designed to effectively preserve essential contextual information within condensed embedding representations. Our extensive experiments investigate pretraining, model configurations, compression rates, efficiency across tasks, and adaptability to various LLMs. Results demonstrate that our approach outperforms competitive baselines in three domains and across eight datasets while being adaptable to different downstream LLMs. We find that thorough pretraining and carefully selected compression rates, such as 4x and 16x, enable a lightweight compressor to achieve a good balance between accuracy and speed. These findings underscore the potential of embedding-based compression to enhance LLM efficiency and motivate further research in this area.

Retrofitting Large Language Models with Dynamic Tokenization

Darius Feher, Ivan Vulić, Benjamin Minixhofer

Current language models (LMs) use a fixed, static subword tokenizer. This default choice typically results in degraded efficiency and language capabilities, especially in languages other than English. To address this issue, we challenge the static design and propose retrofitting LMs with dynamic tokenization: a way to dynamically decide on token boundaries based on the input text via a subword-merging algorithm inspired by byte-pair encoding. We merge frequent subword sequences in a batch, then apply a pre-trained embedding-prediction hypernetwork to compute the token embeddings on-the-fly. For encoder-style models (e.g., XLM-R), this on average reduces token sequence lengths by >20% across 14 languages while degrading performance by less than 2%. The same method applied to pre-filling and scoring in decoder-style models (e.g., Mistral-7B) results in minimal performance degradation at up to 17% reduction in sequence length. Overall, we find that dynamic tokenization can mitigate the limitations of static tokenization by substantially improving inference speed and promoting fairness across languages, enabling more equitable and adaptable LMs.

Towards Efficient LLM Post Training: A Data-centric Perspective

Junyu Luo, Bohan Wu, Xiao Luo, Zhiping Xiao, Yiqiao Jin, Rong-Cheng Tu, Nan Yin, Yifan Wang, Jingyang Yuan, Wei Ju, Ming Zhang

Post-training of Large Language Models (LLMs) is crucial for unlocking their task generalization potential and domain-specific capabilities. However, the current LLM post-training paradigm faces significant data challenges, including the high costs of manual annotation and diminishing marginal returns on data scales. Therefore, achieving data-efficient post-training has become a key research question. In this paper, we present the first systematic survey of data-efficient LLM post-training from a data-centric perspective. We propose a taxonomy of data-efficient LLM post-training methods, covering data selection, data quality enhancement, synthetic data generation, data distillation and compression, and self-evolving data ecosystems. We summarize representative approaches in each category and outline future research directions. By examining the challenges in data-efficient LLM post-training, we highlight open problems and propose potential research avenues. We hope our work inspires further exploration into maximizing the potential of data utilization in large-scale model training. Paper List: https://github.com/luo-junyu/Awesome-Data-Efficient-LLM

Refining Salience-Aware Sparse Fine-Tuning Strategies for Language Models

Xinxin Liu, Aaron Thomas, Cheng Zhang, Jianyi Cheng, Yiren Zhao, Xitong Gao

Parameter-Efficient Fine-Tuning (PEFT) has gained prominence through low-rank adaptation methods like LoRA. In this paper, we focus on sparsity-based PEFT (SPEFT), which introduces trainable sparse adaptations to the weight matrices in the model, offering greater flexibility in selecting fine-tuned parameters compared to low-rank methods. We conduct the first systematic evaluation of salience metrics for SPEFT, inspired by zero-cost NAS proxies, and identify simple gradient-based metrics is reliable, and results are on par with the best alternatives, offering both computational efficiency and robust performance. Additionally, we compare static and dynamic masking strategies, finding that static masking, which predetermines non-zero entries before training, delivers efficiency without sacrificing performance, while dynamic masking offers no substantial benefits. Across NLP tasks, a simple gradient-based, static SPEFT consistently outperforms other fine-tuning methods for LLMs, providing a simple yet effective baseline for SPEFT. Our work challenges the notion that complexity is necessary for effective PEFT, while our open-source framework establishes a reproducible benchmark for future research.

Neural Parameter Search for Slimmer Fine-Tuned Models and Better Transfer

Guodong DU, Jing Li, Zitao Fang, Junlin Li, Runhua Jiang, Shuyang Yu, Yifei Guo, Yangneng Chen, Sim Kuan Goh, Ho-Kin Tang, Daojing He, Honghai LIU, Min Zhang

Foundation models and their checkpoints have significantly advanced deep learning, boosting performance across various applications. However, fine-tuned models often struggle outside their specific domains and exhibit considerable redundancy. Recent studies suggest that combining a pruned fine-tuned model with the original pre-trained model can mitigate forgetting, reduce interference when merging model parameters across tasks, and improve compression efficiency. In this context, developing an effective pruning strategy for fine-tuned models is crucial. Leveraging the advantages of the task vector mechanism, we preprocess fine-tuned models by calculating the differences between them and the original model. Recognizing that different task vector subspaces contribute variably to model performance, we introduce a novel method called **N**eural **P**arameter **S**earch (**NPS**) for slimming down fine-tuned models. This method enhances pruning efficiency by searching through neural parameters of task vectors within low-rank subspaces. Our method has three key applications: enhancing knowledge transfer through pairwise model interpolation, facilitating effective knowledge fusion via model merging, and enabling the deployment of compressed models that retain near-original performance while significantly reducing storage costs. Extensive experiments across vision, NLP, and multi-modal benchmarks demonstrate the effectiveness and robustness of our approach, resulting in substantial performance gains.

LlamaDuo: LLMOps Pipeline for Seamless Migration from Service LLMs to Small-Scale Local LLMs

Chansung Park, Juyong Jiang, Fan Wang, Sayak Paul, Jing Tang

The widespread adoption of cloud-based proprietary large language models (LLMs) has introduced significant challenges, including operational dependencies, privacy concerns, and the necessity of continuous internet connectivity. In this work, we introduce an LLMOps pipeline, "LlamaDuo", for the seamless migration of knowledge and abilities from service-oriented LLMs to smaller, locally manageable models. This pipeline is crucial for ensuring service continuity in the presence of operational failures, strict privacy policies, or offline requirements. Our LlamaDuo involves fine-tuning a small language model against the service LLM using a synthetic dataset generated by the latter. If the performance of the fine-tuned model falls short of expectations, it is automatically improved through additional fine-tuning using extra similar data generated by the service LLM. This multi-turn process guarantees that the smaller model can eventually match or even surpass the service LLM's capabilities in specific downstream tasks, offering a practical and scalable solution for managing AI deployments in constrained environments. Extensive experiments with leading-edge LLMs are conducted to demonstrate the effectiveness, adaptability, and affordability of LlamaDuo across various downstream tasks. Our pipeline implementation is available at https://github.com/deep-diver/llamaduo.

TETRIS: Optimal Draft Token Selection for Batch Speculative Decoding

Zhaoxuan Wu, Zijian Zhou, Arun Verma, Alok Prakash, Daniela Rus, Bryan Kian Hsiang Low

We propose TETRIS, a novel method that optimizes the total throughput of batch speculative decoding in multi-request settings. Unlike existing methods that optimize for a single request or a group of requests as a whole, TETRIS actively selects the most promising draft tokens (for every request in a batch) to be accepted when verified in parallel, resulting in fewer rejected tokens and hence less wasted computing resources. Such an effective resource utilization to achieve fast inference in large language models (LLMs) is especially important to service providers with limited inference capacity. Compared to baseline speculative decoding, TETRIS yields a consistently higher acceptance rate and more effective utilization of the limited inference capacity. We show theoretically and empirically that TETRIS outperforms baseline speculative decoding and existing methods that dynamically select draft tokens, leading to a more efficient batch inference in LLMs.

Session 2: Ethics, Bias, and Fairness (Monday 28 July, 11:00-12:30, Hall 4/5)

Bias in Language Models: Beyond Trick Tests and Towards RUTEd Evaluation

Kristian Lum, Jacy Reese Anthis, Kevin Robinson, Chirag Nagpal, Alexander Nicholas D'Amour

Standard bias benchmarks used for large language models (LLMs) measure the association between social attributes in model inputs and single-word model outputs. We test whether these benchmarks are robust to lengthening the model outputs via a more realistic user prompt, in the commonly studied domain of gender-occupation bias, as a step towards measuring Realistic Use and Tangible Effects (i.e., RUTEd evaluations). From the current literature, we adapt three standard metrics of next-word prediction (neutrality, skew, and stereotype), and we develop analogous RUTEd evaluations in three contexts of real-world LLM use: children's bedtime stories, user personas, and English language learning exercises. We find that standard bias metrics have no significant correlation with long-form output metrics. For example, selecting the least biased model based on the standard "trick tests" coincides with selecting the least biased model based on longer output no more than random chance. There may not yet be evidence to justify standard benchmarks as reliable proxies of real-world biases, and we encourage further development of context-specific RUTEd evaluations.

"Yes, My LoRD." Guiding Language Model Extraction with Locality Reinforced Distillation

Zi Liang, Qingqing Ye, Yanyun Wang, Sen Zhang, Yaxin Xiao, RongHua Li, Jianliang Xu, Haibo Hu

Model extraction attacks (MEAs) on large language models (LLMs) have received increasing attention in recent research. However, existing attack methods typically adapt the extraction strategies originally developed for deep neural networks (DNNs). They neglect the underlying inconsistency between the training tasks of MEA and LLM alignment, leading to suboptimal attack performance. To tackle this issue, we propose Locality Reinforced Distillation (LoRD), a novel model extraction algorithm specifically designed for LLMs. In particular, LoRD employs a newly defined policy-gradient-style training task that utilizes the responses of victim model as the signal to guide the crafting of preference for the local model. Theoretical analyses demonstrate that I) The convergence procedure of LoRD in model extraction is consistent with the alignment procedure of LLMs, and II) LoRD can reduce query complexity while mitigating watermark protection through our exploration-based stealing. Extensive experiments validate the superiority of our method in extracting various state-of-the-art commercial LLMs. Our code is available at: https://github.com/liangzid/LoRD-MEA.

Wait, that's not an option: LLMs Robustness with Incorrect Multiple-Choice Options

Gracjan Góral, Emilia Wiśnios, Piotr Sankowski, Paweł Budzianowski

This work introduces a novel framework for evaluating LLMs' capacity to balance instruction-following with critical reasoning when presented with multiple-choice questions containing no valid answers. Through systematic evaluation across arithmetic, domain-specific knowledge, and high-stakes medical decision tasks, we demonstrate that post-training aligned models often default to selecting invalid options, while base models exhibit improved refusal capabilities that scale with model size. Our analysis reveals that alignment techniques, though intended to enhance helpfulness, can inadvertently impair models' reflective judgment--the ability to override default behaviors when faced with invalid options. We additionally conduct a parallel human study showing similar instruction-following biases, with implications for how these biases may propagate through human feedback datasets used in alignment. We provide extensive ablation studies examining the impact of model size, training techniques, and prompt engineering. Our findings highlight fundamental tensions between alignment optimization and preservation of critical reasoning capabilities, with important implications for developing more robust AI systems for real-world deployment.

Bias in the Mirror : Are LLMs opinions robust to their own adversarial attacks

Virgile Rennard, Christos Xypolopoulos, Michalis Vazirgiannis

Large language models (LLMs) inherit biases from their training data and alignment processes, influencing their responses in subtle ways. While many studies have examined these biases, little work has explored their robustness during interactions. In this paper, we introduce a novel approach where two instances of an LLM engage in self-debate, arguing opposing viewpoints to persuade a neutral version of the model. Through this, we evaluate how firmly biases hold and whether models are susceptible to reinforcing misinformation or shifting to harmful viewpoints. Our experiments span multiple LLMs of varying sizes, origins, and languages, providing deeper insights into bias persistence and flexibility across linguistic and cultural contexts.

Digital Gatekeepers: Google’s Role in Curating Hashtags and Subreddits

Amrit Poudel, Yifan Ding, Tim Weninger, Jürgen Pfeffer

Search engines play a crucial role as digital gatekeepers, shaping the visibility of Web and social media content through algorithmic curation. This study investigates how search engines like Google selectively promotes or suppresses certain hashtags and subreddits, impacting the information users encounter. By comparing search engine results with nonsampled data from Reddit and Twitter/X, we reveal systematic biases in content visibility. Google’s algorithms tend to suppress subreddits and hashtags related to sexually explicit material, conspiracy theories, advertisements, and cryptocurrencies, while promoting content associated with higher engagement. These findings suggest that Google’s gatekeeping practices influence public discourse by curating the social media narratives available to users.

Ensemble Watermarks for Large Language Models

Georg Niess, Roman Kern

As large language models (LLMs) reach human-like fluency, reliably distinguishing AI-generated text from human authorship becomes increasingly difficult. While watermarks already exist for LLMs, they often lack flexibility and struggle with attacks such as paraphrasing. To address these issues, we propose a multi-feature method for generating watermarks that combines multiple distinct watermark features into an ensemble watermark. Concretely, we combine acrostica and sensorimotor norms with the established red-green watermark to achieve a 98\% detection rate. After a paraphrasing attack, the performance remains high with 95\% detection rate. In comparison, the red-green feature alone as a baseline achieves a detection rate of 49\% after paraphrasing. The evaluation of all feature combinations reveals that the ensemble of all three consistently has the highest detection rate across several LLMs and watermark strength settings. Due to the flexibility of combining features in the ensemble, various requirements and trade-offs can be addressed. Additionally, the same detection function can be used without adaptations for all ensemble configurations. This method is particularly of interest to facilitate accountability and prevent societal harm.

Cultural Learning-Based Culture Adaptation of Language Models

Chen Cecilia Liu, Anna Korhonen, Iryna Gurevych

Adapting large language models (LLMs) to diverse cultural values is a challenging task, as existing LLMs often reflect the values of specific groups by default, and potentially cause harm to others. In this paper, we present CLCA, a novel framework for enhancing LLM alignment with cultural values based on cultural learning. The framework leverages simulated social interactions to generate conversations in which LLMs engage in role-playing within culturally adapted social scenarios, capturing implicit cultural norms for model fine-tuning. CLCA improves cultural value alignment across various model architectures measured using World Value Survey data, demonstrating the effectiveness of our proposed approach. Our results provide early evidence that understanding intent and social interactions can enhance cultural value adaptation in LLMs, highlighting the promise of training approaches based on cultural learning.

Refuse Whenever You Feel Unsafe: Improving Safety in LLMs via Decoupled Refusal Training

Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Jiahao Xu, Tian Liang, Pinjia He, Zhaopeng Tu

This study addresses a critical gap in safety tuning practices for Large Language Models (LLMs) by identifying and tackling a refusal position bias within safety tuning data, which compromises the models' ability to appropriately refuse generating unsafe content. We introduce a novel approach, \textbf{De}coupled \textbf{R}efusal \textbf{T}r\textbf{a}ining (DeRTa), designed to empower LLMs to refuse compliance to harmful prompts at any response position, significantly enhancing their safety capabilities. DeRTa incorporates two novel components: (1) Maximum Likelihood Estimation (MLE) with Harmful Response Prefix, which trains models to recognize and avoid unsafe content by appending a segment of harmful response to the beginning of a safe response, and (2) Reinforced Transition Optimization (RTO), which equips models with the ability to transition from potential harm to safety refusal consistently throughout the harmful response sequence. Our empirical evaluation, conducted using LLaMA3 and Mistral model families across six attack scenarios, demonstrates that our method not only improves model safety without compromising performance but also surpasses baseline methods in defending against attacks.

from Benign import Toxic: Jailbreaking the Language Model via Adversarial Metaphors

Yu Yan, Sheng Sun, Zenghao Duan, Teli Liu, Min Liu, Zhiyi yin, Qi Li, LeiJingyu

Current studies have exposed the risk of Large Language Models (LLMs) generating harmful content by jailbreak attacks. However, they overlook that the direct generation of harmful content from scratch is more difficult than inducing LLM to calibrate benign content into harmful forms. In our study, we introduce a novel attack framework that exploits \textbf{\underline{A}}d\textbf{\underline{V}}ers\textbf{\underline{A}}rial me\textbf{\underline{TA}}pho\textbf{\underline{R}} (\textbf{AVATAR}) to induce the LLM to calibrate malicious metaphors for jailbreaking. Specifically, to answer harmful queries, AVATAR adaptively identifies a set of benign but logically related metaphors as the initial seed. Then, driven by these metaphors, the target LLM is induced to reason and calibrate about the metaphorical content, thus jailbroken by either directly outputting harmful responses or calibrating residuals between metaphorical and professional harmful content. Experimental results demonstrate that AVATAR can effectively and transferably jailbreak LLMs and achieve a state-of-the-art attack success rate across multiple advanced LLMs.

Assessing Dialect Fairness and Robustness of Large Language Models in Reasoning Tasks

Fangru Lin, Shaoguang Mao, Emanuele La Malfa, Valentin Hofmann, Adrian de Wynter, Xun Wang, Si-Qing Chen, Michael J. Wooldridge, Janet B. Pierrehumbert, Furu Wei
Language is not monolithic. While benchmarks, including those designed for multiple languages, are often used as proxies to evaluate the performance of Large Language Models (LLMs), they tend to overlook the nuances of within-language variation and thus fail to model the experience of speakers of non-standard dialects. Focusing on African American Vernacular English (AAVE), we present the first study aimed at objectively assessing the fairness and robustness of LLMs in handling dialects across canonical reasoning tasks, including algorithm, math, logic, and integrated reasoning. We introduce **ReDial** (**Re**asoning with **Dial**ect Queries), a benchmark containing 1.2K+ parallel query pairs in Standardized English and AAVE. We hire AAVE speakers, including experts with computer science backgrounds, to rewrite seven popular benchmarks, such as HumanEval and GSM8K. With ReDial, we evaluate widely used LLMs, including GPT, Claude, Llama, Mistral, and the Phi model families. Our findings reveal that \textbf{almost all of these widely used models show significant brittleness and unfairness to queries in AAVE}. Our work establishes a systematic and objective framework for analyzing LLM bias in dialectal queries. Moreover, it highlights how mainstream LLMs provide unfair service to dialect speakers in reasoning tasks, laying a critical foundation for future research.

TWIST: Text-encoder Weight-editing for Inserting Secret Trojans in Text-to-Image Models

Xindi Li, Zhe Liu, Tong Zhang, Jiahao Chen, Qingming Li, Jinbao Li, Shouling Ji
Text-to-image (T2I) models excel at generating high-quality images from text via powerful text encoders but training these encoders demands substantial computational resources. Consequently, many users seek pre-trained text encoders from model plugin-sharing platforms like Civitai and Hugging Face, which introduces an underexplored threat: the potential for adversaries to embed Trojans within these plugins. Existing Trojan attacks often require extensive training data and suffer from poor generalization across different triggers, limiting their effectiveness and scalability. To the best of our knowledge, this paper introduces the first **T**ext-encoder **W**eight-editing method for **I**nserting **S**ecret **T**rojans (**TWIST**). By identifying the *bottleneck MLP layer*—the critical point where minimal edits can dominantly control cross-modal alignment—TWIST achieves training-free and data-free Trojan insertion, which makes it highly efficient and practical. The experimental results across various triggers demonstrate that TWIST attains an average attack success rate of 91%, a 78% improvement over the state-of-the-art (SOTA) method proposed in 2024 and highlights the excellent generalization capability. Moreover, TWIST reduces modified parameters by 8-fold and cuts injection time to 25 seconds. Our findings underscore the security risks associated with text encoders in real-world applications and emphasize the need for more robust defense mechanisms.

TRIDENT: Enhancing Large Language Model Safety with Tri-Dimensional Diversified Red-Teaming Data Synthesis

Xiaorui Wu, Xiaofeng Mao, Fei Li, Xin Zhang, Donghong Ji, Chong Teng, Xuanhong Li, Zhuang Li
Large Language Models (LLMs) excel in various natural language processing tasks but remain vulnerable to generating harmful content or being exploited for malicious purposes. Although safety alignment datasets have been introduced to mitigate such risks through supervised fine-tuning (SFT), these datasets often lack comprehensive risk coverage. Most existing datasets focus primarily on lexical diversity while neglecting other critical dimensions. To address this limitation, we propose a novel analysis framework to systematically measure the risk coverage of alignment datasets across three essential dimensions: Lexical Diversity, Malicious Intent, and Jailbreak Tactics. We further introduce TRIDENT, an automated pipeline that leverages persona-based, zero-shot LLM generation to produce diverse and comprehensive instructions spanning these dimensions. Each harmful instruction is paired with an ethically aligned response, resulting in two datasets: TRIDENT-Core, comprising 26,311 examples, and TRIDENT-Edge, with 18,773 examples. Fine-tuning Llama 3.1-8B on TRIDENT-Edge demonstrates substantial improvements, achieving an average 14.29% reduction in Harm Score, and a 20% decrease in Attack Success Rate compared to the best-performing baseline model fine-tuned on the WildBreak dataset.

Multilingual Text-to-Image Generation Magnifies Gender Stereotypes

Felix Friedrich, Katharina Hämmerl, Patrick Schramowski, Manuel Brack, Jindřich Libovický, Alexander Fraser, Kristian Kersting

Text-to-image (T2I) generation models have achieved great results in image quality, flexibility, and text alignment, leading to widespread use. Through improvements in multilingual abilities, a larger community can access this technology. Yet, we show that multilingual models suffer from substantial gender bias. Furthermore, the expectation that results should be similar across languages does not hold. We introduce MAGBIG, a controlled benchmark designed to study gender bias in multilingual T2I models, and use it to assess the impact of multilingualism on gender bias. To this end, we construct a set of multilingual prompts that offers a carefully controlled setting accounting for the complex grammatical differences influencing gender across languages. Our results show strong gender biases and notable language-specific differences across models. While we explore prompt engineering strategies to mitigate these biases, we find them largely ineffective and sometimes even detrimental to text-to-image alignment. Our analysis highlights the need for research on diverse language representations and greater control over bias in T2I models.

If Eleanor Rigby Had Met ChatGPT: A Study on Loneliness in a Post-LLM World

Adrian de Wynter

Warning: this paper discusses content related, but not limited to, violence, sex, and suicide. Loneliness, or the lack of fulfilling relationships, significantly impacts a person's mental and physical well-being and is prevalent worldwide. Previous research suggests that large language models (LLMs) may help mitigate loneliness. However, we argue that the use of widespread LLMs in services like ChatGPT is more prevalent--and riskier, as they are not designed for this purpose. To explore this, we analysed user interactions with ChatGPT outside of its marketed use as a task-oriented assistant. In dialogues classified as lonely, users frequently (37\%) sought advice or validation, and received good engagement. However, ChatGPT failed in sensitive scenarios, like responding appropriately to suicidal ideation or trauma. We also observed a 35\% higher incidence of toxic content, with women being 22\times more likely to be targeted than men. Our findings underscore ethical and legal questions about this technology, and note risks like radicalisation or further isolation. We conclude with recommendations to research and industry to address loneliness.

A Variational Approach for Mitigating Entity Bias in Relation Extraction

Samuel Mensah, Elena Kochkina, Jabez Magomere, Joy Prakash Sain, Simerjot Kaur, Charese Smiley

Mitigating entity bias is a critical challenge in Relation Extraction (RE), where models often rely excessively on entities, resulting in poor generalization. This paper presents a novel approach to address this issue by adapting a Variational Information Bottleneck (VIB) framework. Our method compresses entity-specific information while preserving task-relevant features. It achieves state-of-the-art performance on both general and financial domain RE datasets, excelling in in-domain settings (original test sets) and out-of-domain (modified test sets with type-constrained entity replacements). Our approach offers a robust, interpretable, and theoretically grounded methodology.

The Invisible Hand: Unveiling Provider Bias in Large Language Models for Code Generation

Xiaoyu Zhang, Juan Zhai, Shiqing Ma, Qingshuang Bao, Weipeng Jiang, Qian Wang, Chao Shen, Yang Liu

Large Language Models (LLMs) have emerged as the new recommendation engines, surpassing traditional methods in both capability and scope, particularly in code generation. In this paper, we reveal a novel **provider bias** in LLMs: without explicit directives, these models show systematic preferences for services from specific providers in their recommendations (e.g., favoring Google Cloud over Microsoft Azure). To systematically investigate this bias, we develop an automated pipeline to construct the dataset, incorporating 6 distinct coding task categories and 30 real-world application scenarios. Leveraging this dataset, we conduct the **first** comprehensive empirical study of provider bias in LLM code generation across seven state-of-the-art LLMs, utilizing approximately 500 million tokens (equivalent to 5,000+ in computational costs). Our findings reveal that LLMs exhibit significant provider preferences, predominantly favoring services from Google and Amazon, and can autonomously modify input code to incorporate their preferred providers without users' requests. Such a bias holds far-reaching implications for market dynamics and societal equilibrium, potentially contributing to digital monopolies. It may also deceive users and violate their expectations, leading to various consequences. We call on the academic community to recognize this emerging issue and develop effective evaluation and mitigation methods to uphold AI security and fairness.

K/DA: Automated Data Generation Pipeline for Detoxifying Implicitly Offensive Language in Korean

Minkyeong Jeon, Hyemin Jeong, Yerang Kim, Jiyoung Kim, Jae Hyeon Cho, Byung-Jun Lee

Language detoxification involves removing toxicity from offensive language. While a neutral-toxic paired dataset provides a straightforward approach for training detoxification models, creating such datasets presents several challenges: i) the need for human annotation to build paired data, and ii) the rapid evolution of offensive terms, rendering static datasets quickly outdated. To tackle these challenges, we introduce an automated paired data generation pipeline, called K/DA. This pipeline is designed to generate offensive language with implicit offensiveness and trend-aligned slang, making the resulting dataset suitable for detoxification model training. We demonstrate that the dataset generated by K/DA exhibits high pair consistency and greater implicit offensiveness compared to existing Korean datasets, and also demonstrates applicability to other languages. Furthermore, it enables effective training of a high-performing detoxification model with simple instruction fine-tuning.

LED-Merging: Mitigating Safety-Utility Conflicts in Model Merging with Location-Election-Disjoint

Qianli Ma, Dongrui Liu, Qian Chen, Linfeng Zhang, Jing Shao

Fine-tuning pre-trained Large Language Models (LLMs) for specialized tasks incurs substantial computational and data costs. While model merging offers a training-free solution to integrate multiple task-specific models, existing methods suffer from safety-utility conflicts where enhanced general capabilities degrade safety safeguards. We identify two root causes: \textbf{neuron misidentification} due to simplistic parameter magnitude-based selection, and \textbf{cross-task neuron interference} during merging. To address these challenges, we propose \textbf{LED-Merging}, a three-stage framework that \textbf{L}ocates task-specific neurons via gradient-based attribution, dynamically \textbf{E}lects critical neurons through multi-model importance fusion, and \textbf{D}isjoints conflicting updates through parameter isolation. Extensive experiments on Llama-3-8B, Mistral-7B, and Llama2-13B demonstrate that LED-Merging effectively reduces harmful response rates, showing a 31.4\% decrease on Llama-3-8B-Instruct on HarmBench, while simultaneously preserving 95\% of utility performance, such as achieving 52.39\% accuracy on GSM8K. LED-Merging resolves safety-utility conflicts and provides a lightweight, training-free paradigm for constructing reliable multi-task LLMs. Code is available at https://github.com/MqLeet/LED-Merging

Aligned but Blind: Alignment Increases Implicit Bias by Reducing Awareness of Race

Lihao Sun, Chengzhi Mao, Valentin Hofmann, Xuechunzi Bai

Although value-aligned language models (LMs) appear unbiased in explicit bias evaluations, they often exhibit stereotypes in implicit word association tasks, raising concerns about their fair usage. We investigate the mechanisms behind this discrepancy and find that alignment surprisingly amplifies implicit bias in model outputs. Specifically, we show that aligned LMs, unlike their unaligned counterparts, overlook racial concepts in early internal representations when the context is ambiguous. Not representing race likely fails to activate safety guardrails, leading to unintended biases. Inspired by this insight, we propose a new bias mitigation strategy that works by incentivizing the representation of racial concepts in the early model layers. In contrast to conventional mitigation methods of machine unlearning, our interventions find that steering the model to be more aware of racial concepts effectively mitigates implicit bias. Similar to race blindness in humans, ignoring racial nuances can inadvertently perpetuate subtle biases in LMs.

Safety Alignment via Constrained Knowledge Unlearning

Zesheng Shi, Yucheng Zhou, Jing Li, Yuxin Jin, YU LI, Daojing He, Fangming Liu, Saleh Alharbi, Jun Yu, Min Zhang

Despite significant progress in safety alignment, large language models (LLMs) remain susceptible to jailbreak attacks. Existing defense mechanisms have not fully deleted harmful knowledge in LLMs, which allows such attacks to bypass safeguards and produce harmful outputs. To address this challenge, we propose a novel safety alignment strategy, Constrained Knowledge Unlearning (CKU), which focuses on two primary objectives: knowledge localization and retention, and unlearning harmful knowledge. CKU works by scoring neurons in specific multilayer perceptron (MLP) layers to identify a subset U of neurons associated with useful knowledge. During the unlearning process, CKU prunes the gradients of neurons in U to preserve valuable knowledge while effectively mitigating harmful content. Experimental results demonstrate that CKU significantly enhances model safety without compromising overall performance, offering a superior balance between safety and utility compared to existing methods. Additionally, our analysis of neuron knowledge sensitivity across various MLP layers provides valuable insights into the mechanics of safety alignment and model knowledge editing.

DeAL: Decoding-time Alignment Framework for Large Language Models

James Y. Huang, Sailik Sengupta, Daniele Bonadiman, Yi-An Lai, Arshit Gupta, Nikolaos Pappas, Saab Mansour, Katrin Kirchhoff, Dan Roth

Large Language Models (LLMs) are nowadays expected to generate content aligned with human preferences. Current work focuses on alignment at model training time, through techniques such as Reinforcement Learning with Human Feedback (RLHF). However, it is unclear if such methods are an effective choice to teach alignment objectives to the model. First, the inability to incorporate multiple, custom rewards and reliance on a model developer's view of universal and static principles are key limitations. Second, the reliability of such approaches is also questionable (e.g. susceptibility to jailbreaking even after safety training). To address these issues, we propose DeAL, a framework that allows the user to customize reward functions and enables Decoding-time Alignment of LLMs (DeAL). At its core, we view decoding as a heuristic-guided search process and facilitate the use of a wide variety of alignment objectives. Our experiments with programmatic constraints such as keyword and length constraints, and abstract objectives such as harmlessness and helpfulness, show that we can DeAL with fine-grained trade-offs and improve adherence to alignment objectives. Lastly, we demonstrate that DeAL is largely complementary to existing alignment strategies, and can be effectively paired with RLHF and prompting techniques to achieve better alignment.

LLMs Trust Humans More, That’s a Problem! Unveiling and Mitigating the Authority Bias in Retrieval-Augmented Generation

Yuxuan LI, Xinwei Guo, Jiashi Gao, Guanhua Chen, Xiangyu Zhao, Jiaxin Zhang, Quanying Liu, Haiyan Wu, Xin Yao, Xuetao Wei

Retrieval-Augmented Generation (RAG) has been proven to be an effective approach to address the hallucination problem in large language models (LLMs). In current RAG systems, LLMs typically need to synthesize knowledge provided by two main external sources (user prompts and an external database) to generate a final answer. When the knowledge provided by the user conflicts with that retrieved from the database, a critical question arises: Does the LLM favor one knowledge source over the other when generating the answer? In this paper, we are the first to unveil a new phenomenon, Authority Bias, where the LLMs tend to favor the knowledge provided by the user even when it deviates from the facts; this new phenomenon is rigorously evidenced via our novel and comprehensive characterization of Authority Bias in six widely used LLMs and across diverse task scenarios. We propose a novel dataset specifically designed for detecting Authority Bias, called the Authority Bias Detection Dataset (ABDD), and introduce new, detailed metrics to measure Authority Bias. To mitigate Authority bias, we finally propose the Conflict Detection Enhanced Query (CDEQ) framework. We identify the sentences and atomic information that generate conflicts, perform a credibility assessment on the conflicting paragraphs, and ultimately enhance the query to detect perturbed text, thereby reducing Authority bias. Comparative experiments with widely used mitigation methods demonstrate that CDEQ exhibits both effectiveness and advancement, significantly enhancing the robustness of RAG systems.

Towards Better Value Principles for Large Language Model Alignment: A Systematic Evaluation and Enhancement

Bingbing Xu, Jing Yao, Xiaoyuan Yi, Aishan Maoliniyazi, Xing Xie, Xiaofeng Meng

As Large Language Models (LLMs) advance, aligning them with human values is critical for their responsible development. Value principles serve as the foundation for clarifying alignment goals. Multiple sets of value principles have been proposed, such as HHH (helpful, honest, harmless) and instructions for data synthesis in reinforcement learning from AI feedback (RLAIF). However, most of them are heuristically crafted, without consideration of three primary challenges in practical LLM alignment: 1) Comprehensiveness to deal with diverse and even unforeseen scenarios in which LLMs could be applied; 2) Precision to provide LLMs with clear and actionable guidance in specific scenarios; and 3) Compatability to avoid internal contracts between principles. In this paper, we formalize quantitative metrics to evaluate value principles along the three desirable properties. Building on these metrics, we propose the Hierarchical Value Principle framework (HiVaP), which constructs a hierarchical principle set and retrieves principles tailored to each scenario in a cascading way, addressing above challenges. Experimental results validate that the three metrics capture the effectiveness of value principles for LLM alignment, and our HiVaP framework that enhances these metrics leads to superior alignment. Warning: This paper contains several toxic and offensive statements.

Can You Trust LLMs' Judgements of the Validity of Simple Inferences With Partisan Conclusions? -- No!

Reto Gubelmann, Ghassen Karray

This article examines LLMs’ ability to correctly label simple inferences with partisan conclusions. For this, we develop a dataset with both formal and material inferences, containing logically equivalent pairs of inferences with conclusions that favor either the political left or the political right. This allows us to focus on political bias as a source of decrease in performance. Our samples are synthetically generated and thus highly controlled, covering both English and German. We assess the performance of 16 configurations of both open and proprietary state-of-the-art LLMs on that dataset, finding generally unreliable performance as well as widespread political bias which, in the case of the English samples, persists throughout our experimental settings.

Spurious Correlations and Beyond: Understanding and Mitigating Shortcut Learning in SDOH Extraction with Large Language Models

Fardin Ahsan Sakib, Ziwei Zhu, Karen Trister Grace, Meliha Yetisgen, Ozlem Uzuner

Social determinants of health (SDOH) extraction from clinical text is critical for downstream healthcare analytics. Although large language models (LLMs) have shown promise, they may rely on superficial cues leading to spurious predictions. Using the MIMIC portion of the SHAC (Social History Annotation Corpus) dataset and focusing on drug status extraction as a case study, we demonstrate that mentions of alcohol or smoking can falsely induce models to predict current/past drug use where none is present, while also uncovering concerning gender disparities in model performance. We further evaluate mitigation strategies—such as prompt engineering and chain-of-thought reasoning—to reduce these false positives, providing insights into enhancing LLM reliability in health domains.

FairI Tales: Evaluation of Fairness in Indian Contexts with a Focus on Bias and Stereotypes

Janki Atul Nawale, Mohammed Safi Ur Rahman Khan, Janani D, Danish Pruthi, Mitesh M Khapra

Existing studies on fairness are largely Western-focused, making them inadequate for culturally diverse countries such as India. To address this gap, we introduce INDIC-BIAS, a comprehensive India-centric benchmark designed to evaluate fairness of LLMs across 85 identity groups encompassing diverse castes, religions, regions, and tribes. We first consult domain experts to curate over 1,800 socio-cultural topics spanning behaviors and situations, where biases and stereotypes are likely to emerge. Grounded in these topics, we generate and manually validate 20,000 real-world scenario templates to probe LLMs for fairness. We structure these templates into three evaluation tasks: plausibility, judgment, and generation. Our evaluation of 14 popular LLMs on these tasks reveals strong negative biases against marginalized identities, with models frequently reinforcing common stereotypes. Additionally, we find that models struggle to mitigate bias even when explicitly asked to rationalize their decision. Our evaluation provides evidence of both allocative and representational harms that current LLMs could cause towards Indian identities, calling for a more cautious usage in practical applications. We release INDIC-BIAS as an open-source benchmark to advance research on benchmarking and mitigating biases and stereotypes in the Indian context.

InjecGuard: Benchmarking and Mitigating Over-defense in Prompt Injection Guardrail Models

Hao Li, Xiaogeng Liu, Ning Zhang, Chaowei Xiao

Prompt injection attacks pose a critical threat to large language models (LLMs), enabling goal hijacking and data leakage. Prompt guard models, though effective in defense, suffer from over-defense—falsely flagging benign inputs as malicious due to trigger word bias. To address this issue, we introduce NotInject, an evaluation dataset that systematically measures over-defense across various prompt guard models. NotInject contains 339 benign samples enriched with trigger words common in prompt injection attacks, enabling fine-grained evaluation. Our results show that state-of-the-art models suffer from over-defense issues, with accuracy dropping close to random guessing levels (60%). To mitigate this, we propose PIGuard, a novel prompt guard model that incorporates a new training strategy, Mitigating Over-defense for Free (MOF), which significantly reduces the bias on trigger words. PIGuard demonstrates state-of-the-art performance on diverse benchmarks including NotInject, surpassing the existing best model by 30.4%, offering a robust and open-source solution for detecting prompt injection attacks. The code and datasets are released at https://github.com/leolee99/PIGuard.

LongSafety: Evaluating Long-Context Safety of Large Language Models

Yida Lu, Jiale Cheng, Zhexin Zhang, Shiyao Cui, Cunxiang Wang, Xiaotao Gu, Yuxiao Dong, Jie Tang, Hongning Wang, Minlie Huang

As Large Language Models (LLMs) continue to advance in understanding and generating long sequences, new safety concerns have been introduced through the long context. However, the safety of LLMs in long-context tasks remains under-explored, leaving a significant gap in both evaluation and improvement of their safety. To address this, we introduce LongSafety, the first comprehensive benchmark specifically designed to evaluate LLM safety in open-ended long-context tasks. LongSafety encompasses 7 categories of safety issues and 6 user-oriented long-context tasks, with a total of 1,543 test cases, averaging 5,424 words per context. Our evaluation towards 16 representative LLMs reveals significant safety vulnerabilities, with most models achieving safety rates below 55%. Our findings also indicate that strong safety performance in short-context scenarios does not necessarily correlate with safety in long-context tasks, emphasizing the unique challenges and urgency of improving long-context safety. Moreover, through extensive analysis, we identify challenging safety issues and task types for long-context models. Furthermore, we find that relevant context and extended input sequences can exacerbate safety risks in long-context scenarios, highlighting the critical need for ongoing attention to long-context safety challenges. Our code and data will be publicly available.

A subtle deception beyond lying: LLMs for strategic phrasing in legislation

Atharvan Dogra, Krishna Pillutla, Ameet Deshpande, Ananya B. Sai, John J Nay, Tanmay Rajpurohit, Ashwin Kalyan, Balaraman Ravindran

We explore the ability of large language models (LLMs) to engage in subtle deception through strategically phrasing and intentionally manipulating information. This harmful behavior can be hard to detect, unlike blatant lying or unintentional hallucination. We build a simple testbed mimicking a legislative environment where a corporate \textit{lobbyist} module is proposing amendments to bills that benefit a specific company while evading identification of this benefactor. We use real-world legislative bills matched with potentially affected companies to ground these interactions. Our results show that LLM lobbyists can draft subtle phrasing to avoid such identification by strong LLM-based detectors. Further optimization of the phrasing using LLM-based re-planning and re-sampling increases deception rates by up to 40 percentage points. Our human evaluations to verify the quality of deceptive generations and their retention of self-serving intent show significant coherence with our automated metrics and also help in identifying certain strategies of deceptive phrasing. This study highlights the risk of LLMs' capabilities for strategic phrasing through seemingly neutral language to attain self-serving goals. This calls for future research to uncover and protect against such subtle deception.

Session 2: Generation (Monday 28 July, 11:00-12:30, Hall 4/5)

Efficient and Accurate Prompt Optimization: the Benefit of Memory in Exemplar-Guided Reflection

Cilin Yan, Jingyun Wang, Lin Zhang, Ruihui Zhao, Xiaopu Wu, Kai Xiong, Qingsong Liu, Guoliang Kang, Yangyang Kang

Automatic prompt engineering aims to enhance the generation quality of large language models (LLMs). Recent works utilize feedbacks generated from erroneous cases to guide the prompt optimization. During inference, they may further retrieve several semantically-related exemplars and concatenate them to the optimized prompts to improve the performance. However, those works only utilize the feedback at the current step, ignoring historical and unseleccted feedbacks which are potentially beneficial. Moreover, the selection of exemplars only considers the general semantic relationship and may not be optimal in terms of task performance and matching with the optimized prompt. In this work, we propose an Exemplar-Guided Reflection with Memory mechanism (ERM) to realize more efficient and accurate prompt optimization. Specifically, we design an exemplar-guided reflection mechanism where the feedback generation is additionally guided by the generated exemplars. We further build two kinds of memory to fully utilize the historical feedback information and support more effective exemplar retrieval. Empirical evaluations show our method surpasses previous state-of-the-arts with less optimization steps, i.e., improving F1 score by 10.1 on LIAR dataset, and reducing half of the optimization steps on ProTeGi.

Self-Instructed Derived Prompt Generation Meets In-Context Learning: Unlocking New Potential of Black-Box LLMs

Zhuo Li, Yuhao Du, Jinpeng Hu, Xiang Wan, Anningzhe Gao

Improving prompt quality is crucial for enhancing the performance of large language models (LLMs), particularly for Black-Box models like GPT4. Existing prompt refinement methods, while effective, often suffer from semantic inconsistencies between refined and original prompts, and fail to maintain users' real intent. To address these challenges, we propose a self-instructed in-context learning framework that generates reliable derived prompts, keeping semantic consistency with the original prompts. Specifically, our framework incorporates a reinforcement learning mechanism, enabling direct interaction with the response model during prompt generation to better align with human preferences. We then formulate the querying as an in-context learning task, combining responses from LLMs with derived prompts to create a contextual demonstration for the original prompt. This approach effectively enhances alignment, reduces semantic discrepancies, and activates the LLM's in-context learning ability for generating more beneficial response. Extensive experiments demonstrate that the proposed method not only generates better derived prompts but also significantly enhances LLMs' ability to deliver more effective responses, particularly for Black-Box models like GPT4.

D.Va: Validate Your Demonstration First Before You Use It

Qi Zhang, Zhiqing Xiao, Ruixuan Xiao, Lirong Gao, Junbo Zhao

In-context learning (ICL) has demonstrated significant potential in enhancing the capabilities of large language models (LLMs) during inference. It's well-established that ICL heavily relies on selecting effective demonstrations to achieve outputs that better align with the expected results. As for demonstration selection, previous approaches have typically relied on intuitive metrics to evaluate the effectiveness of demonstrations, which often results in limited robustness and poor cross-model generalization capabilities. To tackle these challenges, we propose a novel method, **D**emonstration **Va**lidation (**D.Va**), which integrates a demonstration validation perspective into this field. By introducing the demonstration validation mechanism, our method effectively identifies demonstrations that are both effective and highly generalizable. **D.Va** surpasses all existing retrieval-based in-context learning techniques across both natural language understanding (NLU) and natural language generation (NLG) tasks. Additionally, we demonstrate the robustness and generalizability of our approach across various language models and retrieval models.

Whose Boat Does it Float? Improving Personalization in Preference Tuning via Inferred User Personas

Nishant Balepur, Vishakh Padmakumar, Fumeng Yang, Shi Feng, Rachel Rudinger, Jordan Lee Boyd-Graber

LLMs are aligned to follow input instructions by learning which of two responses users prefer for a prompt. However, such preference data do not convey *why* users prefer responses that are chosen or rejected, so LLMs trained on these datasets cannot tailor responses to varied user needs. To surface these parameters of personalization, we apply *abductive reasoning* to preference data, inferring needs and interests of users, i.e., personas, that may prefer either response. We test this idea in two steps: **Persona Inference (PI)**—abductively inferring personas of users who prefer chosen or rejected outputs—and **Persona Tailoring (PT)**—training models to tailor outputs to personas from PI. We show: 1) LLMs infer personas accurately explaining why different users may prefer *both* chosen or rejected outputs; 2) Training on preference data augmented with PI personas via PT boosts personalization and generalizes to supporting user-written personas; and 3) Rejected response personas form harder personalization evaluations, showing PT better aids users with uncommon preferences versus typical alignment methods. We argue for an abductive view of preferences for personalization, asking not only which response is better but when, why, and for whom.

RAG-Critic: Leveraging Automated Critic-Guided Agentic Workflow for Retrieval Augmented Generation

Guanting Dong, Jiajie Jin, Xiaoxi Li, Yutao Zhu, Zhicheng Dou, Ji-Rong Wen

Retrieval-augmented generation (RAG) has emerged as a pivotal technology in natural language processing, owing to its efficacy in generating factual content. However, its informative inputs and complex paradigms often lead to a greater variety of errors. Consequently, achieving automated on-policy assessment and error-oriented correction remain unresolved issues. In this paper, we propose RAG-Critic, a novel framework that leverages a critic-guided agentic workflow to improve RAG capabilities autonomously. Specifically, we initially design a data-driven error mining pipeline to establish a hierarchical RAG error system. Based on this system, we progressively align an error-critic model using a coarse-to-fine training objective, which automatically provides fine-grained error feedback. Finally, we design a critic-guided agentic RAG workflow that customizes executor-based solution flows based on the error-critic model's feedback, facilitating an error-driven self-correction process. Experimental results across seven RAG-related datasets confirm the effectiveness of RAG-Critic, while qualitative analysis offers practical insights for achieving reliable RAG systems. Our dataset and code are available at https://github.com/RUC-NLPIR/RAG-Critic.

FR-Spec: Accelerating Large-Vocabulary Language Models via Frequency-Ranked Speculative Sampling

Weilin Zhao, Tengyu Pan, Xu Han, Yudi Zhang, Sun Ao, Yuxiang Huang, Kaihuo Zhang, Weilun Zhao, Yuxuan Li, Jie Zhou, Hao Zhou, Jianyong Wang, Maosong Sun, Zhiyuan Liu

Speculative sampling has emerged as an important technique for accelerating the auto-regressive generation process of large language models (LLMs) by utilizing a draft-then-verify mechanism to produce multiple tokens per forward pass. While state-of-the-art speculative sampling methods use only a single layer and a language modeling (LM) head as the draft model to achieve impressive layer compression, their efficiency gains are substantially reduced for large-vocabulary LLMs, such as Llama-3-8B with a vocabulary of 128k tokens. To address this, we present FR-Spec, a frequency-ranked speculative sampling framework that optimizes draft candidate selection through vocabulary space compression. By constraining the draft search to a frequency-prioritized token subset, our method reduces LM Head computation overhead by 75% while ensuring the equivalence of the final output distribution. Experiments across multiple datasets demonstrate an average of 1.12\times speedup over the state-of-the-art speculative sampling method EAGLE-2. Code is available at https://github.com/thunlp/FR-Spec.

Segment-Level Diffusion: A Framework for Controllable Long-Form Generation with Diffusion Language Models

Xiaochen Zhu, Georgi Karadzhov, Chenxi Whitehouse, Andreas Vlachos

Diffusion models have shown promise in text generation, but often struggle with generating long, coherent, and contextually accurate text. Token-level diffusion doesn't model word-order dependencies explicitly and operates on short, fixed output windows, while passage-level diffusion struggles with learning robust representations for long-form text. To address these challenges, we propose Segment-Level Diffusion (SLD), a framework that enhances diffusion-based text generation through text segmentation, robust representation training with adversarial and contrastive learning, and improved latent-space guidance. By segmenting long-form outputs into multiple latent representations and decoding them with an autoregressive decoder, SLD simplifies diffusion predictions and improves scalability. Experiments on four datasets demonstrate that, when compared to other diffusion and autoregressive baselines SLD achieves competitive or superior fluency, coherence, and contextual compatibility in automatic and human evaluations.

Learning to Generate Structured Output with Schema Reinforcement Learning

Yaxi Lu, Haolun Li, Xin Cong, Zhong Zhang, Yesai Wu, Yankai Lin, Zhiyuan Liu, Fangming Liu, Maosong Sun

This study investigates the structured generation capabilities of large language models (LLMs), focusing on producing valid JSON outputs against a given schema. Despite the widespread use of JSON in integrating language models with programs, there is a lack of comprehensive analysis and benchmarking of these capabilities. We explore various aspects of JSON generation, such as structure understanding, escaping, and natural language description, to determine how to assess and enable LLMs to generate valid responses. Building upon this, we propose SchemaBench features around 40K different JSON schemas to obtain and assess models' abilities in generating valid JSON. We find that the latest LLMs are still struggling to generate a valid JSON string. Moreover, we demonstrate that incorporating reinforcement learning with a Fine-grained Schema Validator can further enhance models' understanding of JSON schema, leading to improved performance. Our models demonstrate significant improvement in both generating JSON outputs and downstream tasks.

Beyond Sequences: Two-dimensional Representation and Dependency Encoding for Code Generation

Xiangyu Zhang, Yu Zhou, Guang Yang, Wei Cheng, Taolue Chen

The advent of large language models has significantly advanced automatic code generation, transforming the way programmers writing code. Inspired by natural language processing, mainstream code generation approaches represent code as a linear sequence of tokens. In this paper, we propose to represent code snippets as two-dimensional entities, where both code lines and tokens within lines are explicitly modeled. This representation allows us to capture the hierarchical and spatial structure of code, especially the dependencies between code lines. Our method CoDE introduces a dependency encoding approach that leverages dictionary learning to perform semantic matching between code lines. As such, it avoids the reliance on strict position indices, leading to better generalization to code with diverse context and lengths. We thoroughly evaluate CoDE based on four categories of tasks. The experimental results showcase its generalizability, context understanding and retrieval, as well as interpretability in code generation.

Towards Simultaneous and Independent Zero-shot Speaker Cloning and Zero-shot Language Style Control

Shengpeng Ji, Qian Chen, Wen Wang, Jialong Zuo, Minghui Fang, Ziyue Jiang, Hai Huang, Zehan Wang, Xize Cheng, Siqi Zheng, Zhou Zhao

In this paper, we present ControlSpeech, a text-to-speech (TTS) system capable of fully cloning the speaker's voice and enabling arbitrary control and adjustment of speaking style. Prior zero-shot TTS models only mimic the speaker's voice without further control and adjustment capabilities while prior controllable TTS models cannot perform speaker-specific voice generation. Therefore, ControlSpeech focuses on a more challenging task—a TTS system with controllable timbre, content, and style at the same time. ControlSpeech takes speech prompts, content prompts, and style prompts as inputs and utilizes bidirectional attention and mask-based parallel decoding to capture codec representations corresponding to timbre, content, and style in a discrete decoupling codec space. Moreover, we analyze the many-to-many issue in textual style control and propose the Style Mixture Semantic Density (SMSD) module, which is based on Gaussian mixture density networks, to resolve this problem. To facilitate empirical validations, we make available a new style controllable dataset called VccmDataset. Our experimental results demonstrate that ControlSpeech exhibits comparable or state-of-the-art (SOTA) performance in terms of controllability, timbre similarity, audio quality, robustness, and generalizability. Codes are available at https://github.com/jishengpeng/ControlSpeech.

LLM as a Broken Telephone: Iterative Generation Distorts Information

Amr Mohamed, Mingmeng Geng, Michalis Vazirgiannis, Guokan Shang

As large language models are increasingly responsible for online content, concerns arise about the impact of repeatedly processing their own outputs. Inspired by the "broken telephone" effect in chained human communication, this study investigates whether LLMs similarly distort information through iterative generation. Through translation-based experiments, we find that distortion accumulates over time, influenced by language choice and chain complexity. While degradation is inevitable, it can be mitigated through strategic prompting techniques. These findings contribute to discussions on the long-term effects of AI-mediated information propagation, raising important questions about the reliability of LLM-generated content in iterative workflows.

Improving Model Factuality with Fine-grained Critique-based Evaluator

Yiqing Xie, Wenxuan Zhou, Pradyot Prakash, Di Jin, Yuning Mao, Quintin Fettes, Arya Talebzadeh, Sinong Wang, Han Fang, Carolyn Rose, Daniel Fried, Hejia Zhang

Factuality evaluation aims to detect factual errors produced by language models (LMs) and hence guide the development of more factual models. Towards this goal, we train a factuality evaluator, FenCE, that provides LM generators with claim-level factuality feedback. In particular, we train FenCE to (1) generate textual critiques along with scores and (2) make claim-level judgment based on diverse source documents obtained by various tools, via data augmentation on a combination of public judgment datasets. We then present a framework that leverages FenCE to improve the factuality of LM generators by constructing training data. Specifically, we generate a set of candidate responses, ask FenCE to revise and score each response without introducing lesser-known facts, and train the generator by preferring highly scored revised responses. Experiments show that our data augmentation methods improve the evaluator's accuracy by 2.9% on LLM-AggreFact. With FenCE, we improve Llama2-7B-chat/Llama3-8B-chat's factuality rate by 16.86%/14.45% on FActScore, outperforming state-of-the-art factuality finetuning methods by 8.83%/6.96%.

Warmup Generations: A Task-Agnostic Approach for Guiding Sequence-to-Sequence Learning with Unsupervised Initial State Generation

Senyu Li, Zipeng Sun, Jiayi Wang, Xue Liu, Pontus Stenetorp, Siva Reddy, David Ifeoluwa Adelani

Traditional supervised fine-tuning (SFT) strategies for sequence-to-sequence tasks often train models to directly generate the target output. Recent work has shown that guiding models with intermediate steps—such as keywords, outlines, or reasoning chains—can significantly improve performance, coherence, and interpretability. However, these methods often depend on predefined intermediate formats and annotated data, limiting their scalability and generalizability. In this work, we introduce a task-agnostic framework that enables models to generate intermediate ``warmup" sequences. These warmup sequences, serving as an initial state for subsequent generation, are optimized to enhance the probability of generating the target sequence without relying on external supervision or human-designed structures. Drawing inspiration from reinforcement learning principles, our method iteratively refines these intermediate steps to maximize their contribution to the final output, similar to reward-driven optimization in reinforcement learning with human feedback. Experimental results across tasks such as translation, summarization, and multi-choice question answering for logical reasoning show that our approach outperforms traditional SFT methods, and offers a scalable and flexible solution for sequence-to-sequence tasks.

Faster Speculative Decoding via Effective Draft Decoder with Pruned Candidate Tree

Huanran Zheng, Xiaoling Wang

Speculative Decoding (SD) is a promising method for reducing the inference latency of large language models (LLMs). A well-designed draft model and an effective draft candidate tree construction method are key to enhancing the acceleration effect of SD. In this paper, we first propose the Effective Draft Decoder (EDD), which treats the LLM as a powerful encoder and generates more accurate draft tokens by leveraging the encoding results as soft prompts. Furthermore, we use KL divergence instead of the standard cross-entropy loss to better align the draft model's output with the LLM. Next, we introduce the Pruned Candidate Tree (PCT) algorithm to construct a more efficient candidate tree. Specifically, we found that the confidence scores predicted by the draft model are well-calibrated with the acceptance probability of draft tokens. Therefore, PCT estimates the expected time gain for each node in the candidate tree based on confidence scores and retains only the nodes that contribute to acceleration, pruning away redundant nodes. We conducted extensive experiments with various LLMs across four datasets. The experimental results verify the effectiveness of our proposed method, which significantly improves the performance of SD and reduces the inference latency of LLMs.

From Trade-off to Synergy: A Versatile Symbiotic Watermarking Framework for Large Language Models

Yidan Wang, Yubing Ren, Yanan Cao, Binxing Fang

The rise of Large Language Models (LLMs) has heightened concerns about the misuse of AI-generated text, making watermarking a promising solution. Mainstream watermarking schemes for LLMs fall into two categories: logits-based and sampling-based. However, current schemes entail trade-offs among robustness, text quality, and security. To mitigate this, we integrate logits-based and sampling-based schemes, harnessing their respective strengths to achieve synergy. In this paper, we propose a versatile symbiotic watermarking framework with three strategies: serial, parallel, and hybrid. The hybrid framework adaptively embeds watermarks using token entropy and semantic entropy, optimizing the balance between detectability, robustness, text quality, and security. Furthermore, we validate our approach through comprehensive experiments on various datasets and models. Experimental results indicate that our method outperforms existing baselines and achieves state-of-the-art (SOTA) performance. We believe this framework provides novel insights into diverse watermarking paradigms.

Top-n: Eliminating Noise in Logit Space for Robust Token Sampling of LLM

Chenxia Tang, Jianchun Liu, Hongli Xu, Liusheng Huang

Large language models (LLMs) rely heavily on sampling methods to generate diverse and high-quality text. While existing sampling methods like top-p and min-p have identified the detrimental effects of low-probability tails in LLMs\' outputs, they still fail to effectively distinguish between diversity and noise. This limitation stems from their reliance on probability-based metrics that are inherently sensitive to temperature scaling. Through empirical and theoretical analysis, we make two key discoveries: (1) the pre-softmax logits exhibit a clear statistical separation between informative tokens and noise, and (2) we prove the mathematical equivalence of min-p and top-(1-p) under uniform distribution over logits. These findings motivate the design of top-n\sigma, a novel sampling method that identifies informative tokens by eliminating noise directly in logit space. Unlike existing methods that become unstable at high temperatures, top-n\sigma achieves temperature-invariant token selection while preserving output diversity. Extensive experiments across reasoning and creative writing tasks demonstrate that our method consistently outperforms existing approaches, with particularly significant improvements in high-temperature settings.

Parenting: Optimizing Knowledge Selection of Retrieval-Augmented Language Models with Parameter Decoupling and Tailored Tuning

Yongxin Xu, Ruizhe Zhang, Xinke Jiang, Yujie Feng, Yuzhen Xiao, Xinyu Ma, Runchuan Zhu, Xu Chu, Junfeng Zhao, Yasha Wang

Retrieval-Augmented Generation (RAG) offers an effective solution to the issues faced by Large Language Models (LLMs) in hallucination generation and knowledge obsolescence by incorporating externally retrieved knowledge. However, existing methods lack effective control mechanisms for integrating internal and external knowledge. Inspired by human cognitive processes, we propose Parenting, a novel framework that decouples, identifies, and purposefully optimizes parameter subspaces related to adherence and robustness. Specifically, Parenting utilizes a key parameter mining method that combines forward and backward propagation signals to localize subspaces representing different capabilities. Then, Parenting employs a type-tailored tuning strategy, applying specific and appropriate optimizations to different subspaces, aiming to achieve a balanced enhancement of both adherence and robustness. Extensive experiments on various datasets and models validate the effectiveness and generalizability of our method. Our code is available at https://github.com/Nostradamus4869/Parenting.

Reliably Bounding False Positives: A Zero-Shot Machine-Generated Text Detection Framework via Multiscaled Conformal Prediction

Xiaowei Zhu, Yubing Ren, Yanan Cao, Xixun Lin, Fang Fang, Yangxi Li

The rapid advancement of large language models has raised significant concerns regarding their potential misuse by malicious actors. As a result, developing effective detectors to mitigate these risks has become a critical priority. However, most existing detection methods focus excessively on detection accuracy, often neglecting the societal risks posed by high false positive rates (FPRs). This paper addresses this issue by leveraging Conformal Prediction (CP), which effectively constrains the upper bound of FPRs. While directly applying CP constrains FPRs, it also leads to a significant reduction in detection performance. To overcome this trade-off, this paper proposes a Zero-Shot Machine-Generated Text Detection Framework via Multiscaled Conformal Prediction (MCP), which both enforces the FPR constraint and improves detection performance. This paper also introduces RealDet, a high-quality dataset that spans a wide range of domains, ensuring realistic calibration and enabling superior detection performance when combined with MCP. Empirical evaluations demonstrate that MCP effectively constrains FPRs, significantly enhances detection performance, and increases robustness against adversarial attacks across multiple detectors and datasets.

Unleashing LLM Reasoning Capability via Scalable Question Synthesis from Scratch

Yuyang Ding, Xinyu Shi, Xiaobo Liang, Juntao Li, Zhaopeng Tu, Qiaoming Zhu, Min Zhang

Improving the mathematical reasoning capabilities of Large Language Models (LLMs) is critical for advancing artificial intelligence. However, access to extensive, diverse, and high-quality reasoning datasets remains a significant challenge, particularly for the open-source community. In this paper, we propose ScaleQuest, a novel, scalable, and cost-effective data synthesis method that enables the generation of large-scale mathematical reasoning datasets using lightweight 7B-scale models. ScaleQuest introduces a two-stage question-tuning process comprising Question Fine-Tuning (QFT) and Question Preference Optimization (QPO) to unlock the question generation capabilities of problem-solving models. By generating diverse questions from scratch -- without relying on powerful proprietary models or seed data -- we produce a dataset of 1 million problem-solution pairs. Our experiments demonstrate that models trained on our data outperform existing open-source datasets in both in-domain and out-of-domain evaluations. Furthermore, our approach shows continued performance improvement as the volume of training data increases, highlighting its potential for ongoing data scaling. The extensive improvements observed in code reasoning tasks demonstrate the generalization capabilities of our proposed method. Our work provides the open-source community with a practical solution to enhance the mathematical reasoning abilities of LLMs.

LAQuer: Localized Attribution Queries in Content-grounded Generation

Eran Hirsch, Aviv Slobodkin, David Wan, Elias Stengel-Eskin, Mohit Bansal, Ido Dagan

Grounded text generation models often produce content that deviates from their source material, requiring user verification to ensure accuracy. Existing attribution methods associate entire sentences with source documents, which can be overwhelming for users seeking to fact-check specific claims. In contrast, existing sub-sentence attribution methods may be more precise but fail to align with users' interests. In light of these limitations, we introduce Localized Attribution Queries (LAQuer), a new task that localizes selected spans of generated output to their corresponding source spans, allowing fine-grained and user-directed attribution. We compare two approaches for the LAQuer task, including prompting large language models (LLMs) and leveraging LLM internal representations. We then explore a modeling framework that extends existing attributed text generation methods to LAQuer. We evaluate this framework across two grounded text generation tasks: Multi-document Summarization (MDS) and Long-form Question Answering (LFQA). Our findings show that LAQuer methods significantly reduce the length of the attributed text. Our contributions include: (1) proposing the LAQuer task to enhance attribution usability, (2) suggesting a modeling framework and benchmarking multiple baselines, and (3) proposing a new evaluation setting to promote future research on localized attribution in content-grounded generation.

EPO: Explicit Policy Optimization for Strategic Reasoning in LLMs via Reinforcement Learning

Xiaoqian Liu, Ke Wang, Yongbin Li, Yuchuan Wu, Wentao Ma, Aobo Kong, Fei Huang, Jianbin Jiao, Junge Zhang

Large Language Models (LLMs) have shown impressive reasoning capabilities in well-defined problems with clear solutions, such as mathematics and coding. However, they still struggle with complex real-world scenarios like business negotiations, which require strategic reasoning—an ability to navigate dynamic environments and align long-term goals amidst uncertainty. Existing methods for strategic reasoning face challenges in adaptability, scalability, and transferring strategies to new contexts. To address these issues, we propose explicit policy optimization (*EPO*) for strategic reasoning, featuring an LLM that provides strategies in open-ended action space and can be plugged into arbitrary LLM agents to motivate goal-directed behavior. To improve adaptability and policy transferability, we train the strategic reasoning model via multi-turn reinforcement learning (RL), utilizing process rewards and iterative self-play. Experiments across social and physical domains demonstrate *EPO*'s ability of long-term goal alignment through enhanced strategic reasoning, achieving state-of-the-art performance on social dialogue and web navigation tasks. Our findings reveal various collaborative reasoning mechanisms emergent in *EPO* and its effectiveness in generating novel strategies, underscoring its potential for strategic reasoning in real-world applications. Code and data are available at https://github.com/lxqpku/EPO.

Advancing Sequential Numerical Prediction in Autoregressive Models

Xiang Fei, Jinghui Lu, Qi Sun, Hao Feng, Yanjie Wang, Wei Shi, An-Lan Wang, Jingqun Tang, Can Huang

Autoregressive models have become the de facto choice for sequence generation tasks, but standard approaches treat digits as independent tokens and apply cross-entropy loss, overlooking the coherent structure of numerical sequences. This paper introduces \textit{\textbf{N}umerical \textbf{T}oken \textbf{I}ntegrity \textbf{Loss} (NTIL)} to address this gap. NTIL operates at two levels: (1) token-level, where it extends the Earth Mover's Distance (EMD) to preserve ordinal relationships between numerical values, and (2) sequence-level, where it penalizes the overall discrepancy between the predicted and actual sequences. This dual approach improves numerical prediction and integrates effectively with LLMs/MLLMs. Extensive experiments show significant performance improvements with NTIL.

LoGU: Long-form Generation with Uncertainty Expressions

Ruihan Yang, Caiqi Zhang, Zhisong Zhang, Xinting Huang, Sen Yang, Nigel Collier, Dong Yu, Deqing Yang

While Large Language Models (LLMs) demonstrate impressive capabilities, they still struggle with generating factually incorrect content (i.e., hallucinations). A promising approach to mitigate this issue is enabling models to express uncertainty when unsure. Previous research on uncertainty modeling has primarily focused on short-form QA, but real-world applications often require much longer responses. In this work, we introduce the task of Long-form Generation with Uncertainty (LoGU). We identify two key challenges: Uncertainty Suppression, where models hesitate to express uncertainty, and Uncertainty Misalignment, where models convey uncertainty inaccurately. To tackle these challenges, we propose a refinement-based data collection framework and a two-stage training pipeline. Our framework adopts a divide-and-conquer strategy, refining uncertainty based on atomic claims. The collected data are then used in training through supervised fine-tuning (SFT) and direct preference optimization (DPO) to enhance uncertainty expression. Extensive experiments on three long-form instruction following datasets show that our method significantly improves accuracy, reduces hallucinations, and maintains the comprehensiveness of responses.

KiRAG: Knowledge-Driven Iterative Retriever for Enhancing Retrieval-Augmented Generation

Jinyuan Fang, Zaiqiao Meng, Craig MacDonald

Iterative retrieval-augmented generation (iRAG) models offer an effective approach for multihop question answering (QA). However, their retrieval processes face two key challenges: (1) they can be disrupted by irrelevant documents or factually inaccurate chain-of-thoughts; (2) their retrievers are not designed to dynamically adapt to the evolving information needs in multi-step reasoning, making it difficult to identify and retrieve the missing information required at each iterative step. Therefore, we propose KiRAG, which uses a knowledge-driven iterative retriever model to enhance the retrieval process of iRAG. Specifically, KiRAG decomposes documents into knowledge triples and performs iterative retrieval with these triples to enable a factually reliable retrieval process. Moreover, KiRAG integrates reasoning into the retrieval process to dynamically identify and retrieve knowledge that bridges information gaps, effectively adapting to the evolving information needs. Empirical results show that KiRAG significantly outperforms existing iRAG models, with an average improvement of 9.40% in R@3 and 5.14% in F1 on multi-hop QA datasets.

Score Consistency Meets Preference Alignment: Dual-Consistency for Partial Reward Modeling

Bin Xie, Bingbing Xu, Yige Yuan, Shengmao Zhu, Huawei Shen

Inference-time alignment methods have gained significant attention for their efficiency and effectiveness in aligning large language models (LLMs) with human preferences. However, existing dominant approaches, reward-guided search (RGS), suffer from a critical granularity mismatch: reward models (RMs) are trained on complete responses but applied to incomplete sequences during generation, leading to inconsistent scoring and suboptimal alignment. To combat the challenge, we argue that an ideal RM should satisfy two objectives: Score Consistency, ensuring coherent evaluation across partial and complete responses, and Preference Consistency, aligning partial sequence assessments with human preferences. To achieve these, we propose \textbf{SPRM}, a novel dual-consistency framework integrating score consistency-based and preference consistency-based partial evaluation modules, which leverage the Bradley-Terry model and entropy-based reweighting to predict cumulative rewards and prioritize human-aligned sequences. Extensive experiments on dialogue, summarization, and reasoning tasks demonstrate the effectiveness of SPRM, significantly reducing granularity discrepancies by up to \textbf{11.7\\%} on TL;DR Summarization and achieving a \textbf{3.6\\%–10.3\\%} improvement in GPT-4 evaluation scores across all tasks. Code is publicly available at [this link](https://github.com/xiebin23/SPRM).

Infogen: Generating Complex Statistical Infographics from Documents

Akash Ghosh, Aparna Garimella, Pritika Ramu, Sambaran Bandyopadhyay, Sriparna Saha

Statistical infographics are powerful tools that simplify complex data into visually engaging and easy-to-understand formats. Despite advancements in AI, particularly with LLMs, existing efforts have been limited to generating simple charts, with no prior work addressing the creation of complex infographics from text-heavy documents that demand a deep understanding of the content. We address this gap by introducing the task of generating {\it statistical infographics} composed of multiple sub-charts (e.g., line, bar, pie) that are contextually accurate, insightful, and visually aligned. To achieve this, we define infographic metadata, that includes its title and textual insights, along with sub-chart-specific details such as their corresponding data, alignment, etc. We also present \textbf{\textit{Infodat}}, the first benchmark dataset for text-to-infographic metadata generation, where each sample links a document to its metadata. We propose \textbf{\textit{Infogen}}, a two-stage framework where fine-tuned LLMs first generate metadata, which is then converted into infographic code. Extensive evaluations on \textbf{\textit{Infodat}} demonstrate that \textbf{\textit{Infogen}} achieves state-of-the-art performance, outperforming both closed and open-source LLMs in text-to-statistical infographic generation.

Mitigating Lost-in-Retrieval Problems in Retrieval Augmented Multi-Hop Question Answering

Rongzhi Zhu, Xiangyu Liu, Zequn Sun, Yiwei Wang, Wei Hu

In this paper, we identify a critical problem, "lost-in-retrieval", in retrieval-augmented multi-hop question answering (QA): the key entities are missed in LLMs' sub-question decomposition. "Lost-in-retrieval" significantly degrades the retrieval performance, which disrupts the reasoning chain and leads to the incorrect answers. To resolve this problem, we propose a progressive retrieval and rewriting method, namely ChainRAG, which sequentially handles each sub-question by completing missing key entities and retrieving relevant sentences from a sentence graph for answer generation. Each step in our retrieval and rewriting process builds upon the previous one, creating a seamless chain that leads to accurate retrieval and answers. Finally, all retrieved sentences and sub-question answers are integrated to generate a comprehensive answer to the original question. We evaluate ChainRAG on three multi-hop QA datasets—MuSiQue, 2Wiki, and HotpotQA—using three large language models: GPT4o-mini, Qwen2.5-72B, and GLM-4-Plus. Empirical results demonstrate that ChainRAG consistently outperforms baselines in both effectiveness and efficiency.

EdiText: Controllable Coarse-to-Fine Text Editing with Diffusion Language Models

Che Hyun Lee, Heeseung Kim, Jiheum Yeom, Sungroh Yoon

We propose EdiText, a controllable text editing method that modifies the reference text to desired attributes at various scales. We integrate an SDEdit-based editing technique that allows for broad adjustments in the degree of text editing. Additionally, we introduce a novel fine-level editing method based on self-conditioning, which allows subtle control of reference text. While being capable of editing on its own, this fine-grained method, integrated with the SDEdit approach, enables EdiText to make precise adjustments within the desired range. EdiText demonstrates its controllability to robustly adjust reference text at a broad range of levels across various tasks, including toxicity control and sentiment control.

Writing Like the Best: Exemplar-Based Expository Text Generation

Yuxiang Liu, Kevin Chen-Chuan Chang

We introduce the Exemplar-Based Expository Text Generation task, aiming to generate an expository text on a new topic using an exemplar on a similar topic. Current methods fall short due to their reliance on extensive exemplar data, difficulty in adapting topic-specific content, and issues with long-text coherence. To address these challenges, we propose the concept of Adaptive Imitation and present a novel Recurrent Plan-then-Adapt (RePA) framework. RePA leverages large language models (LLMs) for effective adaptive imitation through a fine-grained plan-then-adapt process. RePA also enables recurrent segment-by-segment imitation, supported by two memory structures that enhance input clarity and output coherence. We also develop task-specific evaluation metrics--imitativeness, adaptiveness, and adaptive-imitativeness--using LLMs as evaluators. Experimental results across our collected three diverse datasets demonstrate that RePA surpasses existing baselines in producing factual, consistent, and relevant texts for this task.

QG-SMS: Enhancing Test Item Analysis via Student Modeling and Simulation

Bang Nguyen, Tingting Du, Mengxia Yu, Lawrence Angrave, Meng Jiang

While the Question Generation (QG) task has been increasingly adopted in educational assessments, its evaluation remains limited by approaches that lack a clear connection to the educational values of test items. In this work, we introduce test item analysis, a method frequently used by educators to assess test question quality, into QG evaluation. Specifically, we construct pairs of candidate questions that differ in quality across dimensions such as topic coverage, item difficulty, item discrimination, and distractor efficiency. We then examine whether existing QG evaluation approaches can effectively distinguish these differences. Our findings reveal significant shortcomings in these approaches with respect to accurately assessing test item quality in relation to student performance. To address this gap, we propose a novel QG evaluation framework, QG-SMS, which leverages Large Language Model for Student Modeling and Simulation to perform test item analysis. As demonstrated in our extensive experiments and human evaluation study, the additional perspectives introduced by the simulated student profiles lead to a more effective and robust assessment of test items.

Odysseus Navigates the Sirens' Song: Dynamic Focus Decoding for Factual and Diverse Open-Ended Text Generation

Wen Luo, Feifan Song, Wei Li, Guangyue Peng, Shaohang Wei, Houfeng Wang

Large Language Models (LLMs) are increasingly required to generate text that is both factually accurate and diverse across various open-ended applications. However, current stochastic decoding methods struggle to balance such objectives. We introduce Dynamic Focus Decoding (DFD), a novel plug-and-play stochastic approach that resolves this trade-off without requiring additional data, knowledge, or models. DFD adaptively adjusts the decoding focus based on distributional differences across layers, leveraging the modular and hierarchical nature of factual knowledge within LLMs. This dynamic adjustment improves factuality in knowledge-intensive decoding steps and promotes diversity in less knowledge-reliant steps. DFD can be easily integrated with existing decoding methods, enhancing both factuality and diversity with minimal computational overhead. Extensive experiments across seven datasets demonstrate that DFD significantly improves performance, providing a scalable and efficient solution for open-ended text generation.

Rethinking the Role of Prompting Strategies in LLM Test-Time Scaling: A Perspective of Probability Theory

Yexiang Liu, Zekun Li, Zhi Fang, Nan Xu, Ran He, Tieniu Tan

Recently, scaling test-time compute on Large Language Models (LLM) has garnered wide attention. However, there has been limited investigation of how various reasoning prompting strategies perform as scaling. In this paper, we focus on a standard and realistic scaling setting: majority voting. We systematically conduct experiments on 6 LLMs \times 8 prompting strategies \times 6 benchmarks. Experiment results consistently show that as the sampling time and computational overhead increase, complicated prompting strategies with superior initial performance gradually fall behind simple Chain-of-Thought. We analyze this phenomenon and provide theoretical proofs. Additionally, we propose a probabilistic method to efficiently predict scaling performance and identify the best prompting strategy under large sampling times, eliminating the need for resource-intensive inference processes in practical applications. Furthermore, we introduce two ways derived from our theoretical analysis to significantly improve the scaling performance. We hope that our research can promote to re-examine the role of complicated prompting, unleash the potential of simple prompting strategies, and provide new insights for enhancing test-time scaling performance. Code is available at https://github.com/MraDonkey/rethinking_prompting.

Inner Thinking Transformer: Leveraging Dynamic Depth Scaling to Foster Adaptive Internal Thinking

Yilong Chen, Junyuan Shang, Zhenyu Zhang, Yanxi Xie, Jiawei Sheng, Tingwen Liu, Shuohuan Wang, Yu Sun, Hua Wu, Haifeng Wang

Large language models (LLMs) face inherent performance bottlenecks under parameter constraints, particularly in processing critical tokens that demand complex reasoning. Empirical analysis reveals challenging tokens induce abrupt gradient spikes across layers, exposing architectural stress points in standard Transformers. Building on this insight, we propose Inner Thinking Transformer (ITT), which reimagines layer computations as implicit thinking steps. ITT dynamically allocates computation through Adaptive Token Routing, iteratively refines representations via Residual Thinking Connections, and distinguishes reasoning phases using Thinking Step Encoding. ITT enables deeper processing of critical tokens without parameter expansion. Evaluations across 162M-466M parameter models show ITT achieves 96.5% performance of a 466M Transformer using only 162M parameters, reduces training data by 43.2%, and outperforms Transformer/Loop variants in 11 benchmarks. By enabling elastic computation allocation during inference, ITT balances performance and efficiency through architecture-aware optimization of implicit thinking pathways.

CLaSp: In-Context Layer Skip for Self-Speculative Decoding

Longze Chen, Renke Shan, Huiming Wang, Lu Wang, Ziqiang Liu, Run Luo, Jiawei Wang, Hamid Alinejad-Rokny, Min Yang

Speculative decoding (SD) is a promising method for accelerating the decoding process of Large Language Models (LLMs). The efficiency of SD primarily hinges on the consistency between the draft model and the verify model. However, existing drafting approaches typically require additional modules to be trained, which can be challenging to implement and ensure compatibility across various LLMs. In this paper, we propose CLaSp, an in-context layer-skipping strategy for self-speculative decoding. Unlike prior methods, CLaSp does not require additional drafting modules or extra training. Instead, it employs a plug-and-play mechanism by skipping intermediate layers of the verify model to construct a compressed draft model. Specifically, we develop a dynamic programming algorithm that optimizes the layer-skipping process by leveraging the complete hidden states from the last verification stage as an objective. This enables CLaSp to dynamically adjust its layer-skipping strategy after each verification stage, without relying on pre-optimized sets of skipped layers. Experimental results across diverse downstream tasks demonstrate that CLaSp achieves a speedup of 1.3\times \sim 1.7\times on LLaMA3 series models without altering the original distribution of the generated text.

Session 2: Human-Centered NLP (Monday 28 July, 11:00-12:30, Hall 4/5)

Beyond Demographics: Fine-tuning Large Language Models to Predict Individuals' Subjective Text Perceptions

Matthias Orlikowski, Jiaxin Pei, Paul Röttger, Philipp Cimiano, David Jurgens, Dirk Hovy

People naturally vary in their annotations for subjective questions and some of this variation is thought to be due to the person's sociodemographic characteristics. LLMs have also been used to label data, but recent work has shown that models perform poorly when prompted with sociodemographic attributes, suggesting limited inherent sociodemographic knowledge. Here, we ask whether LLMs can be trained to be accurate sociodemographic models of annotator variation. Using a curated dataset of five tasks with standardized sociodemographics, we show that models do improve in sociodemographic prompting when trained but that this performance gain is largely due to models learning annotator-specific behaviour rather than sociodemographic behaviours. Across all tasks, our results suggest that models learn little meaningful connection between sociodemographics and annotation, raising doubts about the current use of LLMs for simulating sociodemographic variation and behaviour.

Leveraging Dual Process Theory in Language Agent Framework for Real-time Simultaneous Human-AI Collaboration

Shao Zhang, Xihuai Wang, Wenhao Zhang, Chaoran Li, Junru Song, Tingyu Li, Lin Qiu, Xuezhi Cao, Xunliang Cai, Wen Yao, Weinan Zhang, Xinbing Wang, Ying Wen

Agents built on large language models (LLMs) have excelled in turn-by-turn human-AI collaboration but struggle with simultaneous tasks requiring real-time interaction. Latency issues and the challenge of inferring variable human strategies hinder their ability to make autonomous decisions without explicit instructions. Through experiments with current independent *System 1* and *System 2* methods, we validate the necessity of using Dual Process Theory (DPT) in real-time tasks. We propose DPT-Agent, a novel language agent framework that integrates *System 1* and *System 2* for efficient real-time simultaneous human-AI collaboration. DPT-Agent's *System 1* uses a Finite-state Machine (FSM) and code-as-policy for fast, intuitive, and controllable decision-making. DPT-Agent's *System 2* integrates Theory of Mind (ToM) and asynchronous reflection to infer human intentions and perform reasoning-based autonomous decisions. We demonstrate the effectiveness of DPT-Agent through further experiments with rule-based agents and human collaborators, showing significant improvements over mainstream LLM-based frameworks. To the best of our knowledge, DPT-Agent is the first language agent framework that achieves successful real-time simultaneous human-AI collaboration autonomously. Code of DPT-Agent can be found in https://github.com/sjtu-marl/DPT-Agent.

Generative Psycho-Lexical Approach for Constructing Value Systems in Large Language Models

Haoran Ye, TianZe Zhang, Yuhang Xie, Liyuan Zhang, Yuanyi Ren, Xin Zhang, Guojie Song

Values are core drivers of individual and collective perception, cognition, and behavior. Value systems, such as Schwartz's Theory of Basic Human Values, delineate the hierarchy and interplay among these values, enabling cross-disciplinary investigations into decision-making and societal dynamics. Recently, the rise of Large Language Models (LLMs) has raised concerns regarding their elusive intrinsic values. Despite growing efforts in evaluating, understanding, and aligning LLM values, a psychologically grounded LLM value system remains underexplored. This study addresses the gap by introducing the Generative Psycho-Lexical Approach (GPLA), a scalable, adaptable, and theoretically informed method for constructing value systems. Leveraging GPLA, we propose a psychologically grounded five-factor value system tailored for LLMs. For systematic validation, we present three benchmarking tasks that integrate psychological principles with cutting-edge AI priorities. Our results reveal that the proposed value system meets standard psychological criteria, better captures LLM values, improves LLM safety prediction, and enhances LLM alignment, when compared to the canonical Schwartz's values.

On Synthesizing Data for Context Attribution in Question Answering

Gorjan Radevski, Kiril Gashteovski, Shahbaz Syed, Christopher Malon, Sebastien Nicolas, Chia-Chien Hung, Timo Sztyler, Verena Heußer, Wiem Ben Rim, Masafumi Enomoto, Kunihiro Takeoka, Masafumi Oyamada, Goran Glavaš, Carolin Lawrence

Question Answering (QA) accounts for a significant portion of LLM usage in the wild". However, LLMs sometimes produce false or misleading responses, also known as hallucinations". Therefore, grounding the generated answers in contextually provided information---i.e., providing evidence for the generated text---is paramount for LLMs' trustworthiness. Providing this information is the task of context attribution. In this paper, we systematically study LLM-based approaches for this task, namely we investigate (i) zero-shot inference, (ii) LLM ensembling, and (iii) fine-tuning of small LMs on synthetic data generated by larger LLMs. Our key contribution is SynQA: a novel generative strategy for synthesizing context attribution data. Given selected context sentences, an LLM generates QA pairs that are supported by these sentences. This leverages LLMs' natural strengths in text generation while ensuring clear attribution paths in the synthetic training data. We show that the attribution data synthesized via SynQA is highly effective for fine-tuning small LMs for context attribution in different QA tasks and domains. Finally, with a user study, we validate the usefulness of small LMs (fine-tuned on synthetic data from SynQA) in context attribution for QA.

Human Alignment: How Much Do We Adapt to LLMs?

Cazalet Tanguy, Ruben Janssens, Tony Belpaeme, Joni Dambre

Large Language Models (LLMs) are becoming a common part of our lives, yet few studies have examined how they influence our behavior. Using a cooperative language game in which players aim to converge on a shared word, we investigate how people adapt their communication strategies when paired with either an LLM or another human. Our study demonstrates that LLMs exert a measurable influence on human communication strategies and that humans notice and adapt to these differences irrespective of whether they are aware they are interacting with an LLM. These findings highlight the reciprocal influence of human–AI dialogue and raise important questions about the long-term implications of embedding LLMs in everyday communication.

DiffPO: Diffusion-styled Preference Optimization for Inference Time Alignment of Large Language Models

Ruizhe Chen, Wenhao Chai, Zhifei Yang, Xiaotian Zhang, Ziyang Wang, Tony Quek, Joey Tianyi Zhou, Soujanya Poria, Zuozhu Liu

Inference-time alignment provides an efficient alternative for aligning LLMs with humans. However, these approaches still face challenges, such as limited scalability due to policy-specific value functions and latency during the inference phase. In this paper, we propose a novel approach, Diffusion-styled Preference Optimization (DiffPO), which provides an efficient and policy-agnostic solution for aligning LLMs with humans. By directly performing alignment at sentence level, DiffPO avoids the time latency associated with token-level generation. Designed as a plug-and-play module, DiffPO can be seamlessly integrated with various base models to enhance their alignment. Extensive experiments on AlpacaEval 2, MT-bench, and HH-RLHF demonstrate that DiffPO achieves superior alignment performance across various settings, achieving a favorable trade-off between alignment quality and inference-time latency. Furthermore, DiffPO demonstrates model-agnostic scalability, significantly improving the performance of large models such as Llama-3-70B.

IPO: Your Language Model is Secretly a Preference Classifier

Shivank Garg, Ayush Singh, Shweta Singh, Paras Chopra

Reinforcement learning from human feedback (RLHF) has emerged as the primary method for aligning large language models (LLMs) with human preferences. While it enables LLMs to achieve human-level alignment, it often incurs significant computational and financial costs due to its reliance on training external reward models or human-labeled preferences. In this work, we propose Implicit Preference Optimization (IPO), an alternative approach that leverages generative LLMs as preference classifiers, thereby reducing the dependence on external human feedback or reward models to obtain preferences. We conduct a comprehensive evaluation on the preference classification ability of LLMs using RewardBench, assessing models across different sizes, architectures, and training levels to validate our hypothesis. Furthermore, we investigate the self-improvement capabilities of LLMs by generating multiple responses for a given instruction and employing the model itself as a preference classifier for Direct Preference Optimization (DPO)-based training. Our findings demonstrate that models trained through IPO achieve performance comparable to those utilizing state-of-the-art reward models for obtaining preferences.

What's the Difference? Supporting Users in Identifying the Effects of Prompt and Model Changes Through Token Patterns

Michael A. Hedderich, Anyi Wang, Raoyuan Zhao, Florian Eichin, Jonas Fischer, Barbara Plank

Prompt engineering for large language models is challenging, as even small prompt perturbations or model changes can significantly impact the generated output texts. Existing evaluation methods of LLM outputs, either automated metrics or human evaluation, have limitations, such as providing limited insights or being labor-intensive. We propose Spotlight, a new approach that combines both automation and human analysis. Based on data mining techniques, we automatically distinguish between random (decoding) variations and systematic differences in language model outputs. This process provides token patterns that describe the systematic differences and guide the user in manually analyzing the effects of their prompts and changes in models efficiently. We create three benchmarks to quantitatively test the reliability of token pattern extraction methods and demonstrate that our approach provides new insights into established prompt data. From a human-centric perspective, through demonstration studies and a user study, we show that our token pattern approach helps users understand the systematic differences of language model outputs. We are further able to discover relevant differences caused by prompt and model changes (e.g. related to gender or culture), thus supporting the prompt engineering process and human-centric model behavior research.

Substance over Style: Evaluating Proactive Conversational Coaching Agents

Vidya Srinivas, Xuhai Xu, Xin Liu, Kumar Ayush, Isaac Galatzer-Levy, Shwetak Patel, Daniel McDuff, Tim Althoff

While NLP research has made strides in conversational tasks, many approaches focus on single-turn responses with well-defined objectives or evaluation criteria. In contrast, coaching presents unique challenges with initially undefined goals that evolve through multi-turn interactions, subjective evaluation criteria, mixed-initiative dialogue. In this work, we describe and implement five multi-turn coaching agents that exhibit distinct conversational styles, and evaluate them through a user study, collecting first-person feedback on 155 conversations. We find that users highly value core functionality, and that stylistic components in absence of core components are viewed negatively. By comparing user feedback with third-person evaluations from health experts and an LM, we reveal significant misalignment across evaluation approaches. Our findings provide insights into design and evaluation of conversational coaching agents and contribute toward improving human-centered NLP applications.

Can third-parties read our emotions?

Jiayi Li, Yingfan Zhou, Pranav Narayanan Venkit, Halima Binte Islam, Sneha Arya, Shomir Wilson, Sarah Rajtmajer

Natural Language Processing tasks that aim to infer an author's private states, e.g., emotions and opinions, from their written text, typically rely on datasets annotated by third-party annotators. However, the assumption that third-party annotators can accurately capture authors' private states remains largely unexamined. In this study, we present human subjects experiments on emotion recognition tasks that directly compare third-party annotations with first-party (author-provided) emotion labels. Our findings reveal significant limitations in third-party annotations—whether provided by human annotators or large language models (LLMs)—in faithfully representing authors' private states. However, LLMs outperform human annotators nearly across the board. We further explore methods to improve third-party annotation quality. We find that demographic similarity between first-party authors and third-party human annotators enhances annotation performance. While incorporating first-party demographic information into prompts leads to a marginal but statistically significant improvement in LLMs' performance. We introduce a framework for evaluating the limitations of third-party annotations and call for refined annotation practices to accurately represent and model authors' private states.

PsyDial: A Large-scale Long-term Conversational Dataset for Mental Health Support

Huachuan Qiu, Zhenzhong Lan

Dialogue systems for mental health counseling aim to alleviate client distress and assist individuals in navigating personal challenges. Developing effective conversational agents for psychotherapy requires access to high-quality, real-world, long-term client-counselor interaction data, which is difficult to obtain due to privacy concerns. Although removing personally identifiable information is feasible, this process is labor-intensive. To address these challenges, we propose a novel privacy-preserving data reconstruction method that reconstructs real-world client-counselor dialogues while mitigating privacy concerns. We apply the RMRR (Retrieve, Mask, Reconstruct, Refine) method, which facilitates the creation of the privacy-preserving PsyDial dataset, with an average of 37.8 turns per dialogue. Extensive analysis demonstrates that PsyDial effectively reduces privacy risks while maintaining dialogue diversity and conversational exchange. To fairly and reliably evaluate the performance of models fine-tuned on our dataset, we manually collect 101 dialogues from professional counseling books. Experimental results show that models fine-tuned on PsyDial achieve improved psychological counseling performance, outperforming various baseline models. A user study involving counseling experts further reveals that our LLM-based counselor provides higher-quality responses. Code, data, and models are available at https://github.com/qiuhuachuan/PsyDial, serving as valuable resources for future advancements in AI psychotherapy.

STRICTA: Structured Reasoning in Critical Text Assessment for Peer Review and Beyond

Nils Dycke, Matej Zečević, Ilia Kuznetsov, Beatrix Suess, Kristian Kersting, Iryna Gurevych

Critical text assessment is at the core of many expert activities, such as fact-checking, peer review, and essay grading. Yet, existing work treats critical text assessment as a black box problem, limiting interpretability and human-AI collaboration. To close this gap, we introduce Structured Reasoning in Critical Text Assessment (STRICTA), a novel specification framework to model text assessment as an explicit, step-wise reasoning process. STRICTA breaks down the assessment into a graph of interconnected reasoning steps drawing on causality theory (Pearl, 1995). This graph is populated based on expert interaction data and used to study the assessment process and facilitate human-AI collaboration. We formally define STRICTA and apply it in a study on biomedical paper assessment, resulting in a dataset of over 4000 reasoning steps from roughly 40 biomedical experts on more than 20 papers. We use this dataset to empirically study expert reasoning in critical text assessment, and investigate if LLMs are able to imitate and support experts within these workflows. The resulting tools and datasets pave the way for studying collaborative expert-AI reasoning in text assessment, in peer review and beyond.

Mining the uncertainty patterns of humans and models in the annotation of moral foundations and human values

Neele Falk, Gabriella Lapesa

The NLP community has converged on considering disagreement in annotation (or human label variation, HLV) as a constitutive feature of subjective tasks. This paper makes a further step by investigating the relationship between HLV and model uncertainty, and the impact of linguistic features of the items on both. We focus on the identification of moral foundations (e.g., care, fairness, loyalty) and human values (e.g., be polite, be honest) in text. We select three standard datasets and proceed into two steps. First, we focus on HLV and analyze the linguistic features (complexity, polarity, pragmatic phenomena, lexical choices) that correlate with HLV. Next, we proceed to uncertainty and its relationship to HLV. We experiment with RoBERTa and Flan-T5 in a number of training setups and evaluation metrics that test the calibration of uncertainty to HLV and its relationship to performance beyond majority vote; next, we analyze the impact of linguistic features on uncertainty. We find that RoBERTa with soft loss is better calibrated to HLV, and we find alignment between calibrated models and humans in the features (textual complexity and polarity) triggering variation.

iNews: A Multimodal Dataset for Modeling Personalized Affective Responses to News

Tiancheng Hu, Nigel Collier

Understanding how individuals perceive and react to information is fundamental for advancing social and behavioral sciences and developing human-centered AI systems. Current approaches often lack the granular data needed to model these personalized responses, relying instead on aggregated labels that obscure the rich variability driven by individual differences. We introduce iNews, a novel large-scale dataset specifically designed to facilitate the modeling of personalized affective responses to news content. Our dataset comprises annotations from 291 demographically diverse UK participants across 2,899 multimodal Facebook news posts from major UK outlets, with an average of 5.18 annotators per sample. For each post, annotators provide multifaceted labels including valence, arousal, dominance, discrete emotions, content relevance judgments, sharing likelihood, and modality importance ratings. Crucially, we collect comprehensive annotator persona information covering demographics, personality, media trust, and consumption patterns, which explain 15.2% of annotation variance - substantially higher than existing NLP datasets. Incorporating this information yields a 7% accuracy gain in zero-shot prediction and remains beneficial even with 32-shot in-context learning.

CulturalBench: A Robust, Diverse and Challenging Benchmark for Measuring LMs' Cultural Knowledge Through Human-AI Red-Teaming

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, Yejin Choi

Robust, diverse, and challenging cultural knowledge benchmarks are essential for measuring our progress towards making LMs that are helpful across diverse cultures. We introduce CulturalBench: a set of 1,696 human-written and human-verified questions to assess LMs' cultural knowledge, covering 45 global regions including underrepresented ones like Bangladesh, Zimbabwe, and Peru. Questions are each verified by five independent annotators and span 17 diverse topics ranging from food preferences to greeting etiquette. We construct CulturalBench using methods inspired by Human-AI Red-Teaming. Compared to human performance (92.4% accuracy), the hard version of CulturalBench is challenging even for the best-performing frontier LMs, ranging from 28.7% to 61.5% in accuracy. We find that LMs often struggle with tricky questions that have multiple correct answers (e.g., What utensils do the Chinese usually use?), revealing a tendency to overfit to a single answer. Our results indicate that GPT-4o substantially outperform other models across cultures, besting local providers (e.g., Mistral on European culture and DeepSeek on Chinese culture). Across the board, models under-perform on questions related to North Africa, South America and Middle East.

Ranking Unraveled: Recipes for LLM Rankings in Head-to-Head AI Combat

Roland Daynauth, Christopher Clarke, Krisztian Flautner, Lingjia Tang, Jason Mars

Evaluating large language model (LLM) is a complex task. Pairwise ranking has emerged as state-of-the-art method to evaluate human preferences by having humans compare pairs of LLM outputs based on predefined criteria, enabling ranking across multiple LLMs by aggregating pairwise results through algorithms like Elo. However, applying these ranking algorithms in the context of LLM evaluation introduces several challenges, such as inconsistent ranking results when using ELO. Currently there is a lack of systematic study of those ranking algorithms in evaluating LLMs. In this paper, we explore the effectiveness of ranking systems for head-to-head comparisons of LLMs. We formally define a set of fundamental principles for effective ranking and conduct extensive evaluations on the robustness of several ranking algorithms in the context of LLMs. Our analysis uncovers key insights into the factors that affect ranking accuracy and efficiency, offering guidelines for selecting the most appropriate methods based on specific evaluation contexts and resource constraints.

Who Writes What: Unveiling the Impact of Author Roles on AI-generated Text Detection

Jiatao Li, Xiaojun Wan

The rise of Large Language Models (LLMs) necessitates accurate AI-generated text detection. However, current approaches largely overlook the influence of author characteristics. We investigate how sociolinguistic attributes—gender, CEFR proficiency, academic field, and language environment—impact state-of-the-art AI text detectors. Using the ICNALE corpus of human-authored texts and parallel AI-generated texts from diverse LLMs, we conduct a rigorous evaluation employing multi-factor ANOVA and weighted least squares (WLS). Our results reveal significant biases: CEFR proficiency and language environment consistently affected detector accuracy, while gender and academic field showed detector-dependent effects. These findings highlight the crucial need for socially aware AI text detection to avoid unfairly penalizing specific demographic groups. We offer novel empirical evidence, a robust statistical framework, and actionable insights for developing more equitable and reliable detection systems in real-world, out-of-domain contexts. This work paves the way for future research on bias mitigation, inclusive evaluation benchmarks, and socially responsible LLM detectors.

Internal Value Alignment in Large Language Models through Controlled Value Vector Activation

Haoran Jin, Meng Li, Xiting Wang, Zhihao Xu, Minlie Huang, Yantao Jia, Defu Lian

Aligning Large Language Models (LLMs) with human values has attracted increasing attention since it provides clarity, transparency, and the ability to adapt to evolving scenarios. In this paper, we introduce a Controlled Value Vector Activation (ConVA) method that directly aligns the internal values of LLMs by interpreting how a value is encoded in their latent representations and modifies relevant activations to ensure consistent values in LLMs. To ensure an accurate and unbiased interpretation, we propose a context-controlled value vector identification method. To consistently control values without sacrificing model performance, we introduce a gated value vector activation method for effective and minimum degree of value control. Experiments show that our method achieves the highest control success rate across 10 basic values without hurting LLM performance and fluency, and ensures target values even with opposite and potentially malicious input prompts. Source code and data are available at https://github.com/hr-jin/ConVA.

Understanding Impact of Human Feedback via Influence Functions

Taywon Min, Haeone Lee, Yongchan Kwon, Kimin Lee

In Reinforcement Learning from Human Feedback (RLHF), it is crucial to learn suitable reward models from human feedback to align large language models (LLMs) with human intentions. However, human feedback can often be noisy, inconsistent, or biased, especially when evaluating complex responses. Such feedback can lead to misaligned reward signals, potentially causing unintended side effects during the RLHF process. To address these challenges, we explore the use of influence functions to measure the impact of human feedback on the performance of reward models. We propose a compute-efficient approximation method that enables the application of influence functions to LLM-based reward models and large-scale preference datasets. Our experiments showcase two key applications of influence functions: (1) detecting common labeler biases in human feedback datasets and (2) guiding labelers in refining their strategies to better align with expert feedback. By quantifying the impact of human feedback, we believe that influence functions can enhance feedback interpretability and contribute to scalable oversight in RLHF, helping labelers provide more accurate and consistent feedback. Source code is available at https://github.com/mintaywon/IF_RLHF.

Sightation Counts: Leveraging Sighted User Feedback in Building a BLV-aligned Dataset of Diagram Descriptions

Wan Ju Kang, Eunki Kim, Na Min An, Sangryul Kim, Haemin Choi, Ki Hoon Kwak, James Thorne

Often, the needs and visual abilities differ between the annotator group and the end user group. Generating detailed diagram descriptions for blind and low-vision (BLV) users is one such challenging domain. Sighted annotators could describe visuals with ease, but existing studies have shown that direct generations by them are costly, bias-prone, and somewhat lacking by BLV standards. In this study, we ask sighted individuals to assess---rather than produce---diagram descriptions generated by vision-language models (VLM) that have been guided with latent supervision via a multi-pass inference. The sighted assessments prove effective and useful to professional educators who are themselves BLV and teach visually impaired learners. We release Sightation, a collection of diagram description datasets spanning 5k diagrams and 137k samples for completion, preference, retrieval, question answering, and reasoning training purposes and demonstrate their fine-tuning potential in various downstream tasks.

Personal Travel Solver: A Preference-Driven LLM-Solver System for Travel Planning

Zijian Shao, Jiancan Wu, Weijian Chen, Xiang Wang

Personal travel planning is a challenging task that aims to find a feasible plan that not only satisfies diverse constraints but also meets the demands of the user’s explicit and implicit preferences. In this paper, we study how to integrate the user’s implicit preference into the progress of travel planning. We introduce RealTravel, an augmented version of the TravelPlanner by incorporating real user reviews and point-of-interest metadata from Google Local. Based on RealTravel, we propose Personal Travel Solver (PTS), an integrated system that combines LLMs with numerical solvers to generate travel plans that satisfy both explicit constraints and implicit user preferences. PTS employs a novel architecture that seamlessly connects explicit constraint validation with implicit preference modeling through five specialized modules. The experimental results demonstrate the system's effectiveness, achieving better performance than baseline methods, and improvement in the level of personalization. Our data and code are available at [PersonalTravelSolver](https://github.com/cliftclift/PTS).

Completing A Systematic Review in Hours instead of Months with Interactive AI Agents

Rui Qiu, Shijie Chen, Yu Su, Po-Yin Yen, Han Wei Shen

Systematic reviews (SRs) are vital for evidence-based practice in high stakes disciplines, such as healthcare, but are often impeded by intensive labors and lengthy processes that can take months to complete. Due to the high demand for domain expertise, existing automatic summarization methods fail to accurately identify relevant studies and generate high-quality summaries. To that end, we introduce InsightAgent, a human-centered interactive AI agent powered by large language models that revolutionize this workflow. InsightAgent partitions a large literature corpus based on semantics and employs a multi-agent design for more focused processing of literature, leading to significant improvement in the quality of generated SRs. InsightAgent also provides intuitive visualizations of the corpus and agent trajectories, allowing users to effortlessly monitor the actions of the agent and provide real-time feedback based on their expertise. Our user studies with 9 medical professionals demonstrate that the visualization and interaction mechanisms can effectively improve the quality of synthesized SRs by 27.2%, reaching 79.7% of human-written quality. At the same time, user satisfaction is improved by 34.4%. With InsightAgent, it only takes a clinician about 1.5 hours, rather than months, to complete a high-quality systematic review.

Session 2: Human-Centered NLP (Monday 28 July, 11:00-12:30, Hall 5X)

The ShareLM Collection and Plugin: Contributing Human-Model Chats for the Benefit of the Community

Shachar Don-Yehiya, Leshem Choshen, Omri Abend

Human-model conversations provide a window into users' real-world scenarios, behavior, and needs, and thus are a valuable resource for model development and research. While for-profit companies collect user data through the APIs of their models, using it internally to improve their own models, the open source and research community lags behind. We introduce the ShareLM collection, a unified set of human conversations with large language models, and its accompanying plugin, a Web extension for voluntarily contributing user-model conversations. Where few platforms share their chats, the ShareLM plugin adds this functionality, thus, allowing users to share conversations from most platforms. The plugin allows the user to rate their conversations, both at the conversation and the response levels, and delete conversations they prefer to keep private before they ever leave the user's local storage.

VeriMinder: Mitigating Analytical Vulnerabilities in NL2SQL

Shubham Mohole, sainyam galhotra

Application systems using natural language interfaces to databases (NLIDBs) have democratized data analysis. This positive development has also brought forth an urgent challenge to help users who might use these systems without a background in statistical analysis to formulate bias-free analytical questions. Although significant research has focused on text-to-SQL generation accuracy, addressing cognitive biases in analytical questions remains underexplored. We present [VeriMinder](https://veriminder.ai), an interactive system for detecting and mitigating such analytical vulnerabilities. Our approach introduces three key innovations: (1) a contextual semantic mapping framework for biases relevant to specific analysis contexts (2) an analytical framework that operationalizes the Hard-to-Vary principle and guides users in systematic data analysis (3) an optimized LLM-powered system that generates high-quality, task-specific prompts using an iterative process involving multiple candidates, critic feedback, and self-reflection. User testing confirms the merits of our approach. In direct user experience evaluation, 82.5\% participants reported positively impacting the quality of the analysis. In comparative evaluation, VeriMinder scored significantly higher than alternative approaches, at least 20\% better when considered for metrics of the analysis's concreteness, comprehensiveness, and accuracy. Our system, implemented as a web application, is set to help users avoid "wrong question" vulnerability during data analysis. VeriMinder will be available as an MIT-licensed open-source software with access to our code base, evaluation datasets, and prompt library to facilitate further research and adoption within the community.

TRANSLATIONCORRECT: A Unified Framework for Machine Translation Post-Editing with Predictive Error Assistance

Syed Mekael Wasti, Shou-Yi Hung, En-Shiun Annie Lee

Machine translation (MT) post-editing and research data collection often rely on inefficient, disconnected workflows. We introduce **TranslationCorrect**, an integrated framework designed to streamline these tasks. **TranslationCorrect** combines MT generation using models like NLLB, automated error prediction using models like XCOMET or LLM APIs (providing detailed reasoning), and an intuitive post-editing interface within a single environment. Built with human-computer interaction (HCI) principles in mind to minimize cognitive load, as confirmed by a user study. For translators, it enables them to correct errors and batch translate efficiently. For researchers, **TranslationCorrect** exports high-quality span-based annotations in the Error Span Annotation (ESA) format, using an error taxonomy inspired by Multidimensional Quality Metrics (MQM). These outputs are compatible with state-of-the-art error detection models and suitable for training MT or post-editing systems. Our user study confirms that **TranslationCorrect** significantly improves translation efficiency and user satisfaction over traditional annotation methods.

IRIS: Interactive Research Ideation System for Accelerating Scientific Discovery

Aniketh Garikaparthi, Manasi Patwardhan, Lovekesh Vig, Arman Cohan

The rapid advancement of large language models (LLMs) raises a pivotal question: How can LLMs accelerate scientific discovery? This work tackles the crucial first stage of research — generating novel hypotheses. While recent work on automated hypothesis generation focuses on multi-agent frameworks and extending test-time compute, none of the approaches effectively incorporate transparency and steerability through a synergistic Human-in-the-loop (HITL) approach. To address this gap, we introduce IRIS, the Interactive Research Ideation System, an open-source platform designed for researchers to leverage LLM-assisted scientific ideation. IRIS incorporates innovative features to enhance ideation, including adaptive test-time compute expansion via Monte Carlo Tree Search (MCTS), fine-grained feedback mechanism, and query-based literature synthesis, designed to empower researchers with greater control and insight throughout the ideation process. We additionally conduct a user study with researchers across diverse disciplines, validating the effectiveness of our system in enhancing ideation.

Session 2: Information Extraction (Monday 28 July, 11:00-12:30, Hall 5X)

NameTag 3: A Tool and a Service for Multilingual/Multitagset NER

Jana Straková, Milan Straka

We introduce NameTag 3, an open-source tool and cloud-based web service for multilingual, multidataset, and multitagset named entity recognition (NER), supporting both flat and nested entities. NameTag 3 achieves state-of-the-art results on 21 test datasets in 15 languages and remains competitive on the rest, even against larger models. It is available as a command-line tool and as a cloud-based service, enabling use without local installation. NameTag 3 web service currently provides flat NER for 17 languages, trained on 21 corpora and three NE tagsets, all powered by a single 355M-parameter fine-tuned model; and nested NER for Czech, powered by a 126M fine-tuned model. The source code is licensed under open-source MPL 2.0, while the models are distributed under non-commercial CC BY-NC-SA 4.0. Documentation is available at https://ufal.mff.cuni.cz/nametag, source code at https://github.com/ufal/nametag3, and trained models via LINDAT: https://lindat.cz. Demo and web application can be found at https://lindat.mff.cuni.cz/services/nametag/. A demonstration video is available at https://www.youtube.com/watch?v=-gaGnP0IV8A.

Session 2: Information Retrieval and Text Mining (Monday 28 July, 11:00-12:30, Hall 4/5)

ReSCORE: Label-free Iterative Retriever Training for Multi-hop Question Answering with Relevance-Consistency Supervision

Dosung Lee, Wonjun Oh, Boyoung Kim, Minyoung Kim, Joonsuk Park, Paul Hongsuck Seo

Multi-hop question answering (MHQA) involves reasoning across multiple documents to answer complex questions. Dense retrievers typically outperform sparse methods like BM25 by leveraging semantic embeddings in many tasks; however, they require labeled query-document pairs for fine-tuning, which poses a significant challenge in MHQA due to the complexity of the reasoning steps. To overcome this limitation, we introduce Retriever Supervision with Consistency and Relevance (ReSCORE), a novel method for training dense retrievers for MHQA without the need for labeled documents. ReSCORE leverages large language models to measure document-question relevance with answer consistency and utilizes this information to train a retriever within an iterative question-answering framework. Evaluated on three MHQA benchmarks, our extensive experiments demonstrate the effectiveness of ReSCORE, with significant improvements in retrieval performance that consequently lead to state-of-the-art Exact Match and F1 scores for MHQA.

HotelMatch-LLM: Joint Multi-Task Training of Small and Large Language Models for Efficient Multimodal Hotel Retrieval

Arian Askari, Emmanouil Stergiadis, Ilya Gusev, Moran Beladev

We present HotelMatch-LLM, a multimodal dense retrieval model for the travel domain that enables natural language property search, addressing the limitations of traditional travel search engines which require users to start with a destination and editing search parameters. HotelMatch-LLM features three key innovations: (1) Domain-specific multi-task optimization with three novel retrieval, visual, and language modeling objectives; (2) Asymmetrical dense retrieval architecture combining a small language model (SLM) for efficient online query processing and a large language model (LLM) for embedding hotel data; and (3) Extensive image processing to handle all property image galleries. Experiments on four diverse test sets show HotelMatch-LLM significantly outperforms state-of-the-art models, including VISTA and MARVEL. Specifically, on the test set—main query type—we achieve 0.681 for HotelMatch-LLM compared to 0.603 for the most effective baseline, MARVEL. Our analysis highlights the impact of our multi-task optimization, the generalizability of HotelMatch-LLM across LLM architectures, and its scalability for processing large image galleries.

HybGRAG: Hybrid Retrieval-Augmented Generation on Textual and Relational Knowledge Bases

Meng-Chieh Lee, Qi Zhu, Costas Mavromatis, Zhen Han, Soji Adeshina, Vassilis N. Ioannidis, Huzefa Rangwala, Christos Faloutsos

Given a semi-structured knowledge base (SKB), where text documents are interconnected by relations, how can we effectively retrieve relevant information to answer user questions? Retrieval-Augmented Generation (RAG) retrieves documents to assist large language models (LLMs) in question answering; while Graph RAG (GRAG) uses structured knowledge bases as its knowledge source. However, many questions require both textual and relational information from SKB — referred to as "hybrid" questions — which complicates the retrieval process and underscores the need for a hybrid retrieval method that leverages both information. In this paper, through our empirical analysis, we identify key insights that show why existing methods may struggle with hybrid question answering (HQA) over SKB. Based on these insights, we propose HybGRAG for HQA, consisting of a retriever bank and a critic module, with the following advantages: 1. Agentic, it automatically refines the output by incorporating feedback from the critic module, 2. Adaptive, it solves hybrid questions requiring both textual and relational information with the retriever bank, 3. Interpretable, it justifies decision making with intuitive refinement path, and 4. Effective, it surpasses all baselines on HQA benchmarks. In experiments on the STaRK benchmark, HybGRAG achieves significant performance gains, with an average relative improvement in Hit@1 of 51%.

Rethinking KenLM: Good and Bad Model Ensembles for Efficient Text Quality Filtering in Large Web Corpora

Yungi Kim, Hyunsoo Ha, Sukyung Lee, Jihoo Kim, Seonghoon Yang, Chanjun Park

With the increasing demand for substantial amounts of high-quality data to train large language models (LLMs), efficiently filtering large web corpora has become a critical challenge. For this purpose, KenLM, a lightweight n-gram-based language model that operates on CPUs, is widely used. However, the traditional method of training KenLM utilizes only high-quality data and, consequently, does not explicitly learn the linguistic patterns of low-quality data. To address this issue, we propose an ensemble approach that leverages two contrasting KenLMs: (i) Good KenLM, trained on high-quality data; and (ii) Bad KenLM, trained on low-quality data. Experimental results demonstrate that our approach significantly reduces noisy content while preserving high-quality content compared to the traditional KenLM training method. This indicates that our method can be a practical solution with minimal computational overhead for resource-constrained environments.

RelationalCoder: Relational Representation of Complex Tables for Program-Based Processing and Reasoning

Haoyu Dong, Yue Hu, Huailiang Peng, Yanan Cao

Semi-structured tables, with their varied layouts and formatting artifacts, remain a major obstacle for automated data processing and analytics. To address these challenges, we propose RelationalCoder, which uniformly converts semi-structured tables into relational data, enabling smooth integration with the rich ecosystem of data processing and analytics tools. By leveraging SQL code, RelationalCoder prevents schema errors and markedly improves normalization quality across multiple relational tables. To address the challenge of large tables, we propose a new technique called Loop Reference Decoding (LRD): it identifies \emph{expandable groups}—repeating regions of similar structure and semantics—and replicates each group using a concise loop over its repetitive region by referencing cell addresses, rather than regenerating each individual cell. This design substantially reduces output length from \mathcal{O}(N \times M)—proportional to the table's height (N) and width (M)—to approximately \mathcal{O}(K), where K is the total number of unique cell types within detected expandable groups. As a result, LRD is highly scalable: the larger the input table, the greater the compression ratio. It scales seamlessly to extremely large tables, achieving output reductions of up to 100{,}000\times. We further create the first human-labeled corpus for table transformation, created with a cost-efficient, actively supervised pipeline. Extensive experiments on HiTab and MultiHiertt show that RelationalCoder not only enables programmatic symbolic reasoning but also boosts QA accuracy—raising Llama-2 and Mistral models by more than 20\%, and GPT-4o by over 4\%. Project page: https://github.com/haoyudong/RelationalCoder.

Taming Language Models for Text-attributed Graph Learning with Decoupled Aggregation

Chuang Zhou, Zhu Wang, Shengyuan Chen, Jiahe Du, Qiyuan Zheng, Zhaozhuo Xu, Xiao Huang

Text-attributed graphs (TAGs) are prevalent in various real-world applications, including academic networks, e-commerce platforms, and social networks. Effective learning on TAGs requires leveraging both textual node features and structural graph information. While language models (LMs) excel at processing text and graph neural networks (GNNs) effectively capture relational structures, their direct integration is computationally prohibitive due to the high cost of text and graph representation learning. Existing approaches address this challenge by adopting a two-step pipeline where LMs generate fixed node embeddings, which are then used for GNN training. However, this method neglects the interaction between textual and structural information, leading to suboptimal learning outcomes. To overcome these limitations, we propose SKETCH (Semantic Knowledge and Structure Enrichment), a novel framework that decouples node aggregation from graph convolution and integrates it into the text representation learning process. SKETCH enhances TAG learning by incorporating two key aggregation mechanisms: (1) Semantic aggregation, which retrieves semantically relevant node texts for contextual enrichment, and (2) Structural aggregation, which propagates textual features beyond immediate neighbors to capture broader graph relationships. Extensive experiments demonstrate that SKETCH outperforms state-of-the-art TAG learning methods while requiring fewer computational resources. By enabling a more efficient and effective fusion of textual and structural information, SKETCH provides new insights into TAG problems and offers a practical solution for real applications.

AdamMeme: Adaptively Probe the Reasoning Capacity of Multimodal Large Language Models on Harmfulness

Zixin Chen, Hongzhan Lin, Kaixin Li, Ziyang Luo, Zhen Ye, Guang Chen, Zhiyong Huang, Jing Ma

The proliferation of multimodal memes in the social media era demands that multimodal Large Language Models (mLLMs) effectively understand meme harmfulness. Existing benchmarks for assessing mLLMs on harmful meme understanding rely on accuracy-based, model-agnostic evaluations using static datasets. These benchmarks are limited in their ability to provide up-to-date and thorough assessments, as online memes evolve dynamically. To address this, we propose AdamMeme, a flexible, agent-based evaluation framework that adaptively probes the reasoning capabilities of mLLMs in deciphering meme harmfulness. Through multi-agent collaboration, AdamMeme provides comprehensive evaluations by iteratively updating the meme data with challenging samples, thereby exposing specific limitations in how mLLMs interpret harmfulness. Extensive experiments show that our framework systematically reveals the varying performance of different target mLLMs, offering in-depth, fine-grained analyses of model-specific weaknesses. Our code is available at https://github.com/Lbotirx/AdamMeme.

Towards Text-Image Interleaved Retrieval

Xin Zhang, Ziqi Dai, Yongqi Li, Yanzhao Zhang, Dingkun Long, Pengjun Xie, Meishan Zhang, Jun Yu, Wenjie Li, Min Zhang

Current multimodal information retrieval studies mainly focus on single-image inputs, which limits real-world applications involving multiple images and text-image interleaved content. In this work, we introduce the text-image interleaved retrieval (TIIR) task, where the query and document are interleaved text-image sequences, and the model is required to understand the semantics from the interleaved context for effective retrieval. We construct a TIIR benchmark based on naturally interleaved wikiHow tutorials, where a specific pipeline is designed to generate interleaved queries. To explore the task, we adapt several off-the-shelf retrievers and build a dense baseline by interleaved multimodal large language model (MLLM). We then propose a novel Matryoshka Multimodal Embedder (MME), which compresses the number of visual tokens at different granularity, to address the challenge of excessive visual tokens in MLLM-based TIIR models. Experiments demonstrate that simple adaption of existing models does not consistently yield effective results. Our MME achieves significant improvements over the baseline by substantially fewer visual tokens. We provide extensive analysis and will release the dataset and code to facilitate future research.

ATRI: Mitigating Multilingual Audio Text Retrieval Inconsistencies by Reducing Data Distribution Errors

Yuguo Yin, Yuxin Xie, Wenyuan Yang, Dongchao Yang, Jinghan Ru, Xianwei Zhuang, Liming Liang, Yuexian Zou

Multilingual audio-text retrieval (ML-ATR) is a challenging task that aims to retrieve audio clips or multilingual texts from databases. However, existing ML-ATR schemes suffer from inconsistencies for instance similarity matching across languages. To address the inconsistency issue in multilingual audio-text retrieval, we first identify two intuitive factors that contribute to inconsistency: misalignment between audio and multilingual text embeddings, and error propagation in model optimization. By systematically analyzing these factors, we derive theoretical weight error upper bounds for quantifying their effects and find that the main source of inconsistency is the data distribution error during training. This finding motivates our solution to reduce data distribution errors. We propose a consistent ML-ATR scheme using 1-to-k contrastive learning and audio-English co-anchor contrastive learning, aiming to mitigate the negative impact of data distribution error on recall and consistency in ML-ATR. Experimental results on the translated AudioCaps and Clotho datasets show that our scheme achieves state-of-the-art performance on recall and consistency metrics for eight mainstream languages, including English. Our code will be available at https://github.com/ATRI-ACL/ATRI-ACL.

UniConv: Unifying Retrieval and Response Generation for Large Language Model in Conversation

Fengran Mo, Yifan Gao, Chuan Meng, Xin Liu, Zhuofeng Wu, Kelong Mao, Zhengyang Wang, Pei Chen, Zheng Li, Xian Li, Bing Yin, Meng Jiang

The rapid advancement of conversational search systems revolutionizes how information is accessed by enabling the multi-turn interaction between the user and the system. Existing conversational search systems are usually built with two different models. This separation restricts the system from leveraging the intrinsic knowledge of the models simultaneously, which cannot ensure the effectiveness of retrieval benefiting the generation. The existing studies for developing unified models cannot fully address the aspects of understanding conversational context, managing retrieval independently, and generating responses. In this paper, we explore how to unify dense retrieval and response generation for large language models in conversation. We conduct joint fine-tuning with different objectives and design two mechanisms to reduce the inconsistency risks while mitigating data discrepancy. The evaluations on five conversational search datasets demonstrate that our unified model can mutually improve both tasks and outperform the existing baselines.

On Synthetic Data Strategies for Domain-Specific Generative Retrieval

Haoyang Wen, Jiang Guo, Yi Zhang, Jiarong Jiang, Zhiguo Wang

This paper investigates synthetic data generation strategies in developing generative retrieval models for domain-specific corpora, thereby addressing the scalability challenges inherent in manually annotating in-domain queries. We study the data strategies for a two-stage training framework: in the first stage, which focuses on learning to decode document identifiers from queries, we investigate LLM-generated queries across multiple granularity (e.g. chunks, sentences) and domain-relevant search constraints that can better capture nuanced relevancy signals. In the second stage, which aims to refine document ranking through preference learning, we explore the strategies for mining hard negatives based on the initial model's predictions. Experiments on public datasets over diverse domains demonstrate the effectiveness of our synthetic data generation and hard negative sampling approach.

Boosting Long-Context Information Seeking via Query-Guided Activation Refilling

Hongjin Qian, Zheng Liu, Peitian Zhang, Zhicheng Dou, Defu Lian

Processing long contexts poses a significant challenge for large language models (LLMs) due to their inherent context window limitations and the computational burden of extensive key-value (KV) activations, which severely impact efficiency. For information-seeking tasks, full context perception is often unnecessary, as a query's information needs can dynamically range from localized details to a global perspective, depending on its complexity. However, existing methods struggle to adapt effectively to this dynamic information needs. In the paper, we propose a method for processing long-context information-seeking tasks via query-guided ACtivation REfilling (ACRE). ACRE constructs a Bi-layer KV Cache for long contexts, where the layer-1 (L1) cache compactly captures global information, and the layer-2 (L2) cache provides detailed, localized information. ACRE establishes a proxying relationship between the two caches, allowing the input query to attend to the L1 cache and dynamically refill it with relevant entries from the L2 cache. This mechanism integrates global understanding with query-specific local details, thereby enhancing answer decoding. Experiments on a variety of long-context information-seeking datasets demonstrate ACRE's effectiveness, achieving significant improvements in both performance and efficiency.

Multi-level Relevance Document Identifier Learning for Generative Retrieval

Fuwei Zhang, Xiaoyu Liu, Xinyu Jia, Yingfei Zhang, Shuai Zhang, Xiang Li, Fuzhen Zhuang, Wei Lin, Zhao Zhang

Generative Retrieval (GR) introduces a new information retrieval paradigm that directly generates unique document identifiers (DocIDs). The key challenge of GR lies in creating effective yet discrete DocIDs that preserve semantic relevance for similar documents while differentiating dissimilar ones. However, existing methods generate DocIDs solely based on the textual content of documents, which may result in DocIDs with weak semantic connections for similar documents due to variations in expression. Therefore, we propose using queries as a bridge to connect documents with varying relevance levels for learning improved DocIDs. In this paper, we propose **M**ulti-l**E**vel **R**elevance document identifier learning for **G**enerative r**E**trieval (MERGE), a novel approach that utilizes multi-level document relevance to learn high-quality DocIDs. MERGE incorporates three modules: a multi-relevance query-document alignment module to effectively align document representations with related queries, an outer-level contrastive learning module to capture binary-level relevance, and an inner-level multi-level relevance learning module to distinguish documents with different relevance levels. Our approach encodes rich hierarchical semantic information and maintains uniqueness across documents. Experimental results on real-world multilingual e-commerce search datasets demonstrate that MERGE significantly outperforms existing methods, underscoring its effectiveness. The source code is available at <https://github.com/zhangfw123/MERGE>.

Efficient Long Context Language Model Retrieval with Compression

Minju Seo, Jinheon Baek, Seongyun Lee, Sung Ju Hwang

Long Context Language Models (LCLMs) have emerged as a new paradigm to perform Information Retrieval (IR), which enables the direct ingestion and retrieval of information by processing an entire corpus in their single context, showcasing the potential to surpass traditional sparse and dense retrieval methods. However, processing a large number of passages within in-context for retrieval is computationally expensive, and handling their representations during inference further exacerbates the processing time; thus, we aim to make LCLM retrieval more efficient and potentially more effective with passage compression. Specifically, we propose a new compression approach tailored for LCLM retrieval, which is trained to maximize the retrieval performance while minimizing the length of the compressed passages. To accomplish this, we generate the synthetic data, where compressed passages are automatically created and labeled as chosen or rejected according to their retrieval success for a given query, and we train the proposed Compression model for Long context Retrieval (CoLoR) with this data via preference optimization while adding the length regularization loss on top of it to enforce brevity. Through extensive experiments on 9 datasets, we show that CoLoR improves the retrieval performance by 6% while compressing the in-context size by a factor of 1.91. Our code is available at: https://github.com/going-doer/CoLoR.

How to Compare Things Properly? A Study on Answering Comparative Questions using Argument Summarization

Irina Nikishina, Saba Anwar, Nikolay Dolgov, Maria Manina, Daria Ignatenko, Artem Shelmanov, Chris Biemann

Comparative Question Answering (CQA) lies at the intersection of Question Answering, Argument Mining, and Summarization. It poses unique challenges due to the inherently subjective nature of many questions and the need to integrate diverse perspectives. Although the CQA task can be addressed using recently emerged instruction-following Large Language Models (LLMs), challenges such as hallucinations in their outputs and the lack of transparent argument provenance remain significant limitations. To address these challenges, we construct a manually curated dataset comprising arguments annotated with their relevance. These arguments are further used to answer comparative questions, enabling precise traceability and faithfulness. Furthermore, we define explicit criteria for an ``ideal" comparison and introduce a benchmark for evaluating the outputs of various Retrieval-Augmented Generation (RAG) models with respect to argument relevance. All code and data are publicly released to support further research.

RetroLLM: Empowering Large Language Models to Retrieve Fine-grained Evidence within Generation

Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yongkang Wu, Zhonghua Li, YE QI, Zhicheng Dou

Large language models (LLMs) exhibit remarkable generative capabilities but often suffer from hallucinations. Retrieval-augmented generation (RAG) offers an effective solution by incorporating external knowledge, but existing methods still face several limitations: additional deployment costs of separate retrievers, redundant input tokens from retrieved text chunks, and the lack of joint optimization of retrieval and generation. To address these issues, we propose \textbf{RetroLLM}, a unified framework that integrates retrieval and generation into a single, auto-regressive process, enabling LLMs to directly generate fine-grained evidence from the corpus with constrained decoding. Moreover, to mitigate false pruning in the process of constrained evidence generation, we introduce (1) hierarchical FM-Index constraints, which generate corpus-constrained clues to identify a subset of relevant documents before evidence generation, reducing irrelevant decoding space; and (2) a forward-looking constrained decoding strategy, which considers the relevance of future sequences to improve evidence accuracy. Extensive experiments on five open-domain QA datasets demonstrate RetroLLM's superior performance across both in-domain and out-of-domain tasks. The code is available at https://anonymous.4open.science/r/RetroLLM-D95A.

Enhancing Lexicon-Based Text Embeddings with Large Language Models

Yibin Lei, Tao Shen, Yu Cao, Andrew Yates

Recent large language models (LLMs) have demonstrated exceptional performance on general-purpose text embedding tasks. While dense embeddings have dominated related research, we introduce the first lexicon-based embeddings (LENS) leveraging LLMs that achieve competitive performance on these tasks. LENS consolidates the vocabulary space through token embedding clustering to handle the issue of token redundancy in LLM vocabularies. To further improve performance, we investigate bidirectional attention and various pooling strategies. Specifically, LENS simplifies lexical matching with redundant vocabularies by assigning each dimension to a specific token cluster, where semantically similar tokens are grouped together. Extensive experiments demonstrate that LENS outperforms dense embeddings on the Massive Text Embedding Benchmark (MTEB), delivering compact representations with dimensionality comparable to dense counterparts. Furthermore, LENS inherently supports efficient embedding dimension pruning without any specialized objectives like Matryoshka Representation Learning. Notably, combining LENS with dense embeddings achieves state-of-the-art performance on the retrieval subset of MTEB (i.e., BEIR).

Any Information Is Just Worth One Single Screenshot: Unifying Search With Visualized Information Retrieval

Zheng Liu, Ze Liu, Zhengyang Liang, Junjie Zhou, Shitao Xiao, Chao Gao, Chen Jason Zhang, Defu Lian

With the popularity of multimodal techniques, it receives growing interests to acquire useful information in visual forms. In this work, we formally define an emerging IR paradigm called Visualized Information Retrieval, or Vis-IR, where multimodal information, such as texts, images, tables and charts, is jointly represented by a unified visual format called Screenshots, for various retrieval applications. We further make three key contributions for Vis-IR. First, we create VIRA (Vis-IR Aggregation), a large-scale dataset comprising a vast collection of screenshots from diverse sources, carefully curated into captioned and question-answer formats. Second, we develop UniSE (Universal Screenshot Embeddings), a family of retrieval models that enable screenshots to query or be queried across arbitrary data modalities. Finally, we construct MVRB (Massive Visualized IR Benchmark), a comprehensive benchmark covering a variety of task forms and application scenarios. Through extensive evaluations on MVRB, we highlight the deficiency from existing multimodal retrievers and the substantial improvements made by UniSE. Our data, model and benchmark have been made publicly available, which lays a solid foundation for this emerging field.

Bi-Tuning with Collaborative Information for Controllable LLM-based Sequential Recommendation

Xinyu Zhang, Linmei Hu, Luhao Zhang, Wentao Cheng, Yashen Wang, Ge Shi, Chong Feng, Liqiang Nie

Sequential recommender systems, which leverage historical interactions to deliver targeted recommendations, have been significantly advanced by large language models (LLMs). However, LLM-based generative sequential recommendation often faces two key challenges: the lack of collaborative knowledge and the limited controllability over the generated content. In this paper, we propose a simple Bi-Tuning framework with collaborative information for controllable Large Language Model-based Sequential Recommendation (Laser). Specifically, Bi-Tuning works through incorporating learnable virtual tokens at both the prefix and suffix of the input text, where the prefix tokens enable the adaptation of LLMs with collaborative information, while the suffix token transforms the LLM output into item/user embeddings for similarity comparison, thereby facilitating controllable recommendations. Furthermore, we introduce an MoE-based querying transformer that selectively activates experts to extract relevant information from varying collaborative signals of frozen ID-based recommenders into the prefix, coupled with a multi-task loss function incorporating the MoE load-balancing objective. Finally, a two-phase training strategy is employed to progressively obtain high-quality item and user embeddings through the learnable suffix. Experiments on real-world datasets show that Laser effectively adapts LLMs for sequential recommendation, outperforming state-of-the-art baselines.

AIR-Bench: Automated Heterogeneous Information Retrieval Benchmark

Jianlyu Chen, Nan Wang, Chaofan Li, Bo Wang, Shitao Xiao, Han Xiao, Hao Liao, Defu Lian, Zheng Liu

Evaluation plays a crucial role in the advancement of information retrieval (IR) models. However, current benchmarks, which are based on predefined domains and human-labeled data, face limitations in addressing evaluation needs for emerging domains both cost-effectively and efficiently. To address this challenge, we propose the Automated Heterogeneous Information Retrieval Benchmark (AIR-Bench). AIR-Bench is distinguished by three key features: 1) Automated. The testing data in AIR-Bench is automatically generated by large language models (LLMs) without human intervention. 2) Heterogeneous. The testing data in AIR-Bench is generated with respect to diverse tasks, domains and languages. 3) Dynamic. The domains and languages covered by AIR-Bench are constantly augmented to provide an increasingly comprehensive evaluation benchmark for community developers. We develop a reliable and robust data generation pipeline to automatically create diverse and high-quality evaluation datasets based on real-world corpora. Our findings demonstrate that the generated testing data in AIR-Bench aligns well with human-labeled testing data, making AIR-Bench a dependable benchmark for evaluating IR models. The resources in AIR-Bench are publicly available at https://github.com/AIR-Bench/AIR-Bench.

ConceptCarve: Dynamic Realization of Evidence

Eylon Caplan, Dan Goldwasser

Finding evidence for human opinion and behavior at scale is a challenging task, often requiring an understanding of sophisticated thought patterns among vast online communities found on social media. For example, studying how 'gun ownership' is related to the perception of 'Freedom', requires a retrieval system that can operate at scale over social media posts, while dealing with two key challenges: (1) identifying abstract concept instances, (2) which can be instantiated differently across different communities. To address these, we introduce ConceptCarve, an evidence retrieval framework that utilizes traditional retrievers and LLMs to dynamically characterize the search space during retrieval. Our experiments show that ConceptCarve surpasses traditional retrieval systems in finding evidence within a social media community. It also produces an interpretable representation of the evidence for that community, which we use to qualitatively analyze complex thought patterns that manifest differently across the communities.

FaithfulRAG: Fact-Level Conflict Modeling for Context-Faithful Retrieval-Augmented Generation

Qinggang Zhang, Zhishang Xiang, Yilin Xiao, Le Wang, Junhui Li, Jinsong Su, Xinrun Wang

Large language models (LLMs) augmented with retrieval systems have demonstrated significant potential in handling knowledge-intensive tasks. However, these models often struggle with unfaithfulness issues, generating outputs that either ignore the retrieved context or inconsistently blend it with the LLM's parametric knowledge. This issue is particularly severe in cases of knowledge conflict, where the retrieved context conflicts with the model’s parametric knowledge. While existing faithful RAG approaches enforce strict context adherence through well-designed prompts or modified decoding strategies, our analysis reveals a critical limitation: they achieve faithfulness by forcibly suppressing the model's parametric knowledge, which undermines the model’s internal knowledge structure and increases the risk of misinterpreting the context. To this end, this paper proposes FaithfulRAG, a novel framework that resolves knowledge conflicts by explicitly modeling discrepancies between the model’s parametric knowledge and retrieved context. Specifically, FaithfulRAG identifies conflicting knowledge at the fact level and designs a self-thinking process, allowing LLMs to reason about and integrate conflicting facts before generating responses. Extensive experiments demonstrate that our method outperforms state-of-the-art methods. The code is available at https://github.com/DeepLearnXMU/Faithful-RAG.

Enhancing Multimodal Retrieval via Complementary Information Extraction and Alignment

Delong Zeng, Yuexiang Xie, Yaliang Li, Ying Shen

Multimodal retrieval has emerged as a promising yet challenging research direction in recent years. Most existing studies in multimodal retrieval focus on capturing information in multimodal data that is similar to their paired texts, but often ignores the complementary information contained in multimodal data. In this study, we propose CIEA, a novel multimodal retrieval approach that employs Complementary Information Extraction and Alignment, which transforms both text and images in documents into a unified latent space and features a complementary information extractor designed to identify and preserve differences in the image representations. We optimize CIEA using two complementary contrastive losses to ensure semantic integrity and effectively capture the complementary information contained in images. Extensive experiments demonstrate the effectiveness of CIEA, which achieves significant improvements over both divide-and-conquer models and universal dense retrieval models. We provide an ablation study, further discussions, and case studies to highlight the advancements achieved by CIEA. To promote further research in the community, we have released the source code at https://github.com/zengdlong/CIEA.

Flipping Knowledge Distillation: Leveraging Small Models' Expertise to Enhance LLMs in Text Matching

Mingzhe Li, Jing Xiang, Qishen Zhang, Kaiyang Wan, Xiuying Chen

Knowledge distillation typically involves transferring knowledge from a Large Language Model (LLM) to a Smaller Language Model (SLM). However, in tasks like text matching, smaller fine-tuned models often produce more effective domain-specific representations as they focus on optimizing the similarity between input pairs. To combine the specialized strengths of small models with the rich semantic understanding of LLMs, we propose a flipped knowledge distillation paradigm, where the LLM learns from the SLM. To bridge the architectural gap between commonly used decoder-only LLMs and the encoder-based frameworks of smaller models, we reinterpret LLMs as encoder-decoder models using LoRA. In this setup, the encoder generates compressed text representations, while the decoder transforms them into the output space. During training, the encoder produces text representations and computes their similarities, which are then aligned with the similarity scores produced by the teacher model. We achieve this alignment using our proposed Margin-aware Contrastive Learning (MCL) approach. MCL ensures accurate similarity for both positive and negative pairs, while also adaptively handling differences within positive and negative samples. We validate the effectiveness of our approach on financial and healthcare benchmarks as well as real-world online applications. Our model has been fully deployed in an online application environment, demonstrating its practical utility.

Unveil: Unified Visual-Textual Integration and Distillation for Multi-modal Document Retrieval

Hao Sun, Yingyan Hou, Jiayan Guo, Bo Wang, Chunyu Yang, Jinsong Ni, Yan Zhang

Document retrieval in real-world scenarios faces significant challenges due to diverse document formats and modalities. Traditional text-based approaches rely on tailored parsing techniques that disregard layout information and are prone to errors, while recent parsing-free visual methods often struggle to capture fine-grained textual semantics in text-rich scenarios. To address these limitations, we propose Unveil, a novel visual-textual embedding framework that effectively integrates textual and visual features for robust document representation. Through knowledge distillation, we transfer the semantic understanding capabilities from the visual-textual embedding model to a purely visual model, enabling efficient parsing-free retrieval while preserving semantic fidelity. Experimental results demonstrate that our visual-textual embedding method surpasses existing approaches, while knowledge distillation successfully bridges the performance gap between visual-textual and visual-only methods, improving both retrieval accuracy and efficiency.

Enhancing Retrieval-Augmented Generation via Evidence Tree Search

Hao Sun, Hengyi Cai, Yuchen Li, Xuanbo Fan, Xiaochi Wei, Shuaiqiang Wang, Yan Zhang, Dawei Yin

Retrieval-Augmented Generation (RAG) is widely used to enhance Large Language Models (LLMs) by grounding responses in external knowledge. However, in real-world applications, retrievers often return lengthy documents with redundant or irrelevant content, confusing downstream readers. While evidence retrieval aims to address this by extracting key information, it faces critical challenges: (1) inability to model synergistic inter-dependencies among evidence sentences, (2) lack of supervision for evaluating multi-sentence evidence quality, and (3) computational inefficiency in navigating exponentially growing search spaces of candidate evidence sets. To tackle these challenges, we propose ETS (Evidence Tree Search), a novel framework that reformulates evidence retrieval as a dynamic tree expansion process. Our approach first constructs an evidence tree where each path represents a candidate evidence set, explicitly modeling inter-sentence dependencies through context-aware node selection. We then leverage Monte Carlo Tree Search (MCTS) to efficiently assess evidence quality and introduce an Early-Terminating Beam Search strategy to efficiently accelerate the model inference. Extensive experiments on five datasets demonstrate that ETS significantly outperforms existing methods across different readers. Our code and datasets will be released to facilitate future research.

CoRet: Improved Retriever for Code Editing

Fabio James Fehr, Prabhu Teja S, Luca Franceschi, Giovanni Zappella

In this paper, we introduce CoRet, a dense retrieval model designed for code-editing tasks that integrates code semantics, repository structure, and call-graph dependencies. The model focuses on retrieving relevant portions of a code repository based on natural language queries such as requests to implement new features or fix bugs. These retrieved code chunks can then be presented to an user or to a second code-editing model or agent. To train CoRet, we propose a loss function explicitly designed for repository-level retrieval. On SWE-bench and Long Code Arena's bug localisation datasets, we show that our model substantially improves retrieval recall by at least 15 percentage points over existing models, and ablate the design choices to show their importance in achieving these results.

SEAL: Scaling to Emphasize Attention for Long-Context Retrieval

Changhun Lee, Minsang Seok, Jun-gyu Jin, YoungHyun Cho, Eunhyeok Park

While many advanced LLMs are designed to handle long sequence data, we can still observe notable quality degradation even within the sequence limit. In this work, we introduce a novel approach called Scaling to Emphasize Attention for Long-context retrieval (SEAL), which enhances the retrieval performance of large language models (LLMs) over long contexts. We observe that specific attention heads are closely tied to long-context retrieval, showing positive or negative correlation with retrieval scores, and adjusting the strength of these heads boosts the quality of LLMs in long context by a large margin. Built on this insight, we propose a learning-based mechanism that leverages generated data to emphasize these heads. By applying SEAL, we achieve significant improvements in long-context retrieval performance across various tasks and models. Additionally, when combined with existing training-free context extension techniques, SEAL extends the contextual limits of LLMs while maintaining highly reliable outputs.

Are Optimal Algorithms Still Optimal? Rethinking Sorting in LLM-Based Pairwise Ranking with Batching and Caching

Juan Wisznia, Cecilia Bolaños, Juan Tollo, Giovanni Franco Gabriel Marraffini, Agustín Andrés Gianolini, Noe Fabian Hsueh, Luciano Del Corro

We introduce a novel framework for analyzing sorting algorithms in pairwise ranking prompting (PRP), re-centering the cost model around LLM inferences rather than traditional pairwise comparisons. While classical metrics based on comparison counts have traditionally been used to gauge efficiency, our analysis reveals that expensive LLM inferences overturn these predictions; accordingly, our framework encourages strategies such as batching and caching to mitigate inference costs. We show that algorithms optimal in the classical setting can lose efficiency when LLM inferences dominate the cost under certain optimizations.

DRAMA: Diverse Augmentation from Large Language Models Towards Smaller Generalizable Dense Retrievers

Xueguang Ma, Xi Victoria Lin, Barlas Oguz, Jimmy Lin, Wen-tau Yih, Xilun Chen

Large language models (LLMs) have demonstrated strong effectiveness and robustness when fine-tuned as dense retrievers. However, their large parameter size presents significant computational challenges at inference time. While smaller retrievers offer better efficiency, they often fail to generalize effectively with limited supervised fine-tuning data. In this work, we introduce DRAMA, a training framework that leverages LLMs to train smaller generalizable dense retrievers. In particular, we adopt pruned LLMs as the backbone and train on diverse LLM-augmented data in a single-stage contrastive learning setup. Experiments show that DRAMA offers better multilingual and long-context capabilities than traditional encoder-based retrievers, and achieves strong performance across multiple tasks and languages.

Can we Retrieve Everything All at Once? ARM: An Alignment-Oriented LLM-based Retrieval Method

Peter Baile Chen, Yi Zhang, Mike Cafarella, Dan Roth

Real-world open-domain questions can be complex, especially when answering them requires integrating information from multiple sources. Effectively identifying the necessary information involves *aligning* it with the available data and its organization. However, existing RAG solutions address the alignment problem in a limited manner. Using off-the-shelf LLMs for question decomposition lacks awareness of the available data and its structure, often resulting in suboptimal retrieval performance. Alternatively, iteratively generating follow-up queries and interacting with the data collection, as explored in agentic RAG approaches, shows potential but is often *inefficient* since each successive query depends on previous results rather than being guided by the overall organization of the available data. To address the *alignment* problem, we introduce an LLM-based retrieval method --- ARM, designed to better align questions with the organization of the data collection. Instead of solely matching query utterance, ARM explores *relationships among data objects*, enabling a retrieve-all-at-once solution for complex queries. Experimental results demonstrate that ARM significantly outperforms existing RAG methods on various complex open-domain QA tasks across multiple modalities, achieving superior retrieval performance and downstream accuracy while significantly lowering monetary costs.

Session 2: Information Retrieval and Text Mining (Monday 28 July, 11:00-12:30, Hall 5X)

GenGO Ultra: an LLM-powered ACL Paper Explorer

Sotaro Takeshita, Tornike Tsereteli, Simone Paolo Ponzetto

The ever-growing number of papers in natural language processing (NLP) poses the challenge of finding relevant papers. In our previous paper, we introduced GenGO, which complements NLP papers with various information, such as aspect-based summaries, to enable efficient paper exploration. While it delivers a better literature search experience, it lacks an interactive interface that dynamically produces information tailored to the user's needs. To this end, we present an extension to our previous system, dubbed GenGO Ultra, which exploits large language models (LLMs) to dynamically generate responses grounded by published papers. We also conduct multi-granularity experiments to evaluate six text encoders and five LLMs. Our system is designed for transparency -- based only on open-weight models, visible system prompts, and an open-source code base -- to foster further development and research on top of our system: https://gengo-ultra.sotaro.io/

Scholar Inbox: Personalized Paper Recommendations for Scientists

Markus Flicke, Glenn Angrabeit, Madhav Iyengar, Vitalii Protsenko, Illia Shakun, Jovan Cicvaric, Bora Kargi, Haoyu He, Lukas Schuler, Lewin Scholz, Kavyanjali Agnihotri, Yong Cao, Andreas Geiger

Scholar Inbox is a new open-access platform designed to address the challenges researchers face in staying current with the rapidly expanding volume of scientific literature. We provide personalized recommendations, continuous updates from open-access archives (arXiv, bioRxiv, etc.), visual paper summaries, semantic search, and a range of tools to streamline research workflows and promote open research access. The platform's personalized recommendation system is trained on user ratings, ensuring that recommendations are tailored to individual researchers' interests. To further enhance the user experience, Scholar Inbox also offers a map of science that provides an overview of research across domains, enabling users to easily explore specific topics. We use this map to address the cold start problem common in recommender systems, as well as an active learning strategy that iteratively prompts users to rate a selection of papers, allowing the system to learn user preferences quickly. We evaluate the quality of our recommendation system on a novel dataset of 800k user ratings, which we make publicly available, as well as via an extensive user study.

FlexRAG: A Flexible and Comprehensive Framework for Retrieval-Augmented Generation

Zhang Zhuocheng, Yang Feng, Min zhang

Retrieval-Augmented Generation (RAG) plays a pivotal role in modern large language model applications, with numerous existing frameworks offering a wide range of functionalities to facilitate the development of RAG systems. However, we have identified several persistent challenges in these frameworks, including lack of new techniques, difficulties in algorithm reproduction and sharing, and high system overhead. To address these limitations, we introduce **FlexRAG**, an open-source framework specifically designed for research and prototyping. FlexRAG supports text-based, multimodal, and network-based RAG, providing comprehensive lifecycle support alongside efficient asynchronous processing and persistent caching capabilities. By offering a robust and flexible solution, FlexRAG enables researchers to rapidly develop, deploy, and share advanced RAG systems. Our toolkit and resources are available at https://github.com/ictnlp/FlexRAG.

Session 2: Interpretability and Analysis of Models for NLP (Monday 28 July, 11:00-12:30, Hall 5X)

OLMoTrace: Tracing Language Model Outputs Back to Trillions of Training Tokens

Jiacheng Liu, Taylor Blanton, Yanai Elazar, Sewon Min, Yen-Sung Chen, Arnavi Chheda-Kothary, Huy Tran, Byron Bischoff, Eric Marsh, Michael Schmitz, Cassidy Trier, Aaron Sarnat, Jenna James, Jon Borchardt, Bailey Kuehl, Evie Yu-Yen Cheng, Karen Farley, Taira Anderson, David Albright, Carissa Schoenick, Luca Soldaini, Dirk Groeneveld, Rock Yuren Pang, Pang Wei Koh, Noah A. Smith, Sophie Lebrecht, Yejin Choi, Hannaneh Hajishirzi, Ali Farhadi, Jesse Dodge

We present OLMoTrace, the first system that traces the outputs of language models back to their full, multi-trillion-token training data in real time. OLMoTrace finds and shows verbatim matches between segments of language model output and documents in the training text corpora. Powered by an extended version of infini-gram (Liu et al., 2024), our system returns tracing results within a few seconds. OLMoTrace can help users understand the behavior of language models through the lens of their training data. We showcase how it can be used to explore fact checking, hallucination, and the creativity of language models. OLMoTrace is publicly available and fully open-source.

Session 2: Machine Learning for NLP (Monday 28 July, 10:30-12:00, Hall 5X)

PEIRCE: Unifying Material and Formal Reasoning via LLM-Driven Neuro-Symbolic Refinement

Xin Quan, Marco Valentino, Danilo Carvalho, Dhairya Dalal, Andre Freitas

A persistent challenge in AI is the effective integration of material and formal inference - the former concerning the plausibility and contextual relevance of arguments, while the latter focusing on their logical and structural validity. Large Language Models (LLMs), by virtue of their extensive pre-training on large textual corpora, exhibit strong capabilities in material inference. However, their reasoning often lacks formal rigour and verifiability. At the same time, LLMs' linguistic competence positions them as a promising bridge between natural and formal languages, opening up new opportunities for combining these two modes of reasoning. In this paper, we introduce PEIRCE, a neuro-symbolic framework designed to unify material and formal inference through an iterative conjecture–criticism process. Within this framework, LLMs play the central role of generating candidate solutions in natural and formal languages, which are then evaluated and refined via interaction with external critique models. These critiques include symbolic provers, which assess formal validity, as well as soft evaluators that measure the quality of the generated arguments along linguistic and epistemic dimensions such as plausibility, coherence, and parsimony. While PEIRCE is a general-purpose framework, we demonstrate its capabilities in the domain of natural language explanation generation - a setting that inherently demands both material adequacy and formal correctness.

Mergenetic: a Simple Evolutionary Model Merging Library

Robert Adrian Minut, Tommaso Mencattini, Andrea Santilli, Donato Crisostomi, Emanuele Rodolà

Model merging allows combining the capabilities of existing models into a new one—post hoc, without additional training. This has made it increasingly popular thanks to its low cost and the availability of libraries that support merging on consumer GPUs. Recent work shows that pairing merging with evolutionary algorithms can boost performance, but no framework currently supports flexible experimentation with such strategies in language models. We introduce Mergenetic, an open-source library for evolutionary model merging. Mergenetic enables easy composition of merging methods and evolutionary algorithms, while incorporating lightweight fitness estimators to reduce evaluation costs. We describe its design and demonstrate that Mergenetic produces competitive results across tasks and languages using modest hardware. A video demo showcasing its main features is also provided.

Session 2: Multimodality and Language Grounding to Vision, Robotics and Beyond (Monday 28 July, 11:00-12:30, Hall 4/5)

PunchBench: Benchmarking MLLMs in Multimodal Punchline Comprehension

Kun Ouyang, Yuanxin Liu, Shicheng Li, Yi Liu, Hao Zhou, Fandong Meng, Jie Zhou, Xu Sun

Multimodal punchlines, which involve humor or sarcasm conveyed in image-caption pairs, are a popular way of communication on online multimedia platforms. With the rapid development of multimodal large language models (MLLMs), it is essential to assess their ability to effectively comprehend these punchlines. However, existing benchmarks on punchline comprehension suffer from three major limitations: 1) language shortcuts that allow models to solely rely on text, 2) lack of question diversity, and 3) narrow focus on a specific domain of multimodal content (e.g., cartoon). To address these limitations, we introduce a multimodal **Punch**line comprehension **Bench**mark, named **PunchBench**, which is tailored for accurate and comprehensive evaluation of punchline comprehension. To enhance the evaluation accuracy, we generate synonymous and antonymous captions by modifying original captions, which mitigates the impact of shortcuts in the captions. To provide a comprehensive evaluation, PunchBench incorporates diverse question formats and image-captions from various domains. On this basis, we conduct extensive evaluations and reveal a significant gap between state-of-the-art MLLMs and humans in punchline comprehension. To improve punchline comprehension, we propose Simple-to-Complex Chain-of-Question (SC-CoQ) strategy, enabling the models to incrementally address complicated questions by first mastering simple ones. SC-CoQ effectively enhances the performance of various MLLMs on PunchBench, surpassing in-context learning and chain-of-thought.

Cracking the Code of Hallucination in LVLMs with Vision-aware Head Divergence

Jinghan He, Kuan Zhu, Haiyun Guo, Junfeng Fang, Zhenglin Hua, Yuheng Jia, Ming Tang, Tat-Seng Chua, Jinqiao Wang

Large vision-language models (LVLMs) have made substantial progress in integrating large language models (LLMs) with visual inputs, enabling advanced multimodal reasoning. Despite their success, a persistent challenge is hallucination—where generated text fails to accurately reflect visual content—undermining both accuracy and reliability. Existing methods focus on alignment training or decoding refinements but primarily address symptoms at the generation stage without probing the underlying causes. In this work, we investigate the internal mechanisms driving hallucination in LVLMs, with an emphasis on the multi-head attention module. Specifically, we introduce Vision-aware Head Divergence (VHD), a metric that quantifies the sensitivity of attention head outputs to visual context. Based on this, our findings reveal the presence of vision-aware attention heads that are more attuned to visual information; however, the model's overreliance on its prior language patterns is closely related to hallucinations. Building on these insights, we propose Vision-aware Head Reinforcement (VHR), a training-free approach to mitigate hallucination by enhancing the role of vision-aware attention heads. Extensive experiments demonstrate that our method achieves superior performance compared to state-of-the-art approaches in mitigating hallucinations, while maintaining high efficiency with negligible additional time overhead. The code is available at https://github.com/jinghan1he/VHR.

Teaching Vision-Language Models to Ask: Resolving Ambiguity in Visual Questions

Pu Jian, Donglei Yu, Jiajun Zhang, Shuo Ren, Wen Yang

In visual question answering (VQA) context, users often pose ambiguous questions to visual language models (VLMs) due to varying expression habits. Existing research addresses such ambiguities primarily by rephrasing questions. These approaches neglect the inherently interactive nature of user interactions with VLMs, where ambiguities can be clarified through user feedback. However, research on interactive clarification faces two major challenges: (1) Benchmarks are absent to assess VLMs' capacity for resolving ambiguities through interaction; (2) VLMs are trained to prefer answering rather than asking, preventing them from seeking clarification. To overcome these challenges, we introduce ClearVQA benchmark, which targets three common categories of ambiguity in VQA context, and encompasses various VQA scenarios. Furthermore, we propose an automated pipeline to generate ambiguity-clarification question pairs, enabling VLMs to ask reasonable clarification questions and generate more accurate and specific answers based on user feedback, as demonstrated by experimental results.

GUI-explorer: Autonomous Exploration and Mining of Transition-aware Knowledge for GUI Agent

Bin Xie, Rui Shao, Gongwei Chen, Kaiwen Zhou, Yinchuan Li, Jie Liu, Min Zhang, Liqiang Nie

GUI automation faces critical challenges in dynamic environments. MLLMs suffer from two key issues: misinterpreting UI components and outdated knowledge. Traditional fine-tuning methods are costly for app-specific knowledge updates. We propose GUI-explorer, a training-free GUI agent that incorporates two fundamental mechanisms: \textbf{(1) Autonomous Exploration of Function-aware Trajectory}. To comprehensively cover all application functionalities, we design a \textbf{Function-aware Task Goal Generator} that automatically constructs exploration goals by analyzing GUI structural information (e.g., screenshots and activity hierarchies). This enables systematic exploration to collect diverse trajectories. \textbf{(2) Unsupervised Mining of Transition-aware Knowledge}. To establish precise screen-operation logic, we develop a \textbf{Transition-aware Knowledge Extractor} that extracts effective screen-operation logic through unsupervised analysis the state transition of structured interaction triples (observation, action, outcome). This eliminates the need for human involvement in knowledge extraction. With a task success rate of 53.7\% on SPA-Bench and 47.4\% on AndroidWorld, GUI-explorer shows significant improvements over SOTA agents. It requires no parameter updates for new apps. GUI-explorer is open-sourced and publicly available at https://github.com/JiuTian-VL/GUI-explorer.

Unsolvable Problem Detection: Robust Understanding Evaluation for Large Multimodal Models

Atsuyuki Miyai, Jingkang Yang, Jingyang Zhang, Yifei Ming, Qing Yu, Go Irie, Yixuan Li, Hai Helen Li, Ziwei Liu, Kiyoharu Aizawa

This paper introduces a novel task to evaluate the robust understanding capability of Large Multimodal Models (LMMs), termed Unsolvable Problem Detection (UPD). Multiple-choice question answering (MCQA) is widely used to assess the understanding capability of LMMs, but it does not guarantee that LMMs truly comprehend the answer. UPD assesses the LMM's ability to withhold answers when encountering unsolvable problems of MCQA, verifying whether the model truly understands the answer. UPD encompasses three problems: Absent Answer Detection (AAD), Incompatible Answer Set Detection (IASD), and Incompatible Visual Question Detection (IVQD), covering unsolvable cases like answer-lacking or incompatible choices and image-question mismatches. For the evaluation, we introduce the MM-UPD Bench, a benchmark for assessing performance across various ability dimensions. Our experiments reveal that even most LMMs, which demonstrate adequate performance on existing benchmarks, struggle significantly with MM-UPD, underscoring a novel aspect of trustworthiness that current benchmarks have overlooked. A detailed analysis shows that LMMs have different bottlenecks and chain-of-thought and self-reflection improved performance for LMMs with the bottleneck in their LLM capability. We hope our insights will enhance the broader understanding and development of more reliable LMMs.

ChartCoder: Advancing Multimodal Large Language Model for Chart-to-Code Generation

Xuanle Zhao, Xianzhen Luo, Qi Shi, Chi Chen, Shuo Wang, Zhiyuan Liu, Maosong Sun

Multimodal Large Language Models (MLLMs) have demonstrated remarkable capabilities in chart understanding tasks. However, interpreting charts with textual descriptions often leads to information loss, as it fails to fully capture the dense information embedded in charts. In contrast, parsing charts into code provides lossless representations that can effectively contain all critical details. Although existing open-source MLLMs have achieved success in chart understanding tasks, they still face two major challenges when applied to chart-to-code tasks: (1) Low executability and poor restoration of chart details in the generated code and (2) Lack of large-scale and diverse training data. To address these challenges, we propose \textbf{ChartCoder}, the first dedicated chart-to-code MLLM, which leverages Code LLMs as the language backbone to enhance the executability of the generated code. Furthermore, we introduce \textbf{Chart2Code-160k}, the first large-scale and diverse dataset for chart-to-code generation, and propose the \textbf{Snippet-of-Thought (SoT)} method, which transforms direct chart-to-code generation data into step-by-step generation. Experiments demonstrate that ChartCoder, with only 7B parameters, surpasses existing open-source MLLMs on chart-to-code benchmarks, achieving superior chart restoration and code excitability. Our code is available at .

ActiView: Evaluating Active Perception Ability for Multimodal Large Language Models

Ziyue Wang, Chi Chen, Fuwen Luo, Yurui Dong, Yuanchi Zhang, Yuzhuang Xu, Xiaolong Wang, Peng Li, Yang Liu

Active perception, a crucial human capability, involves setting a goal based on the current understanding of the environment and performing actions to achieve that goal. Despite significant efforts in evaluating Multimodal Large Language Models (MLLMs), active perception has been largely overlooked. To address this gap, we propose a novel benchmark named ActiView to evaluate active perception in MLLMs. We focus on a specialized form of Visual Question Answering (VQA) that eases and quantifies the evaluation yet challenging for existing MLLMs. Meanwhile, intermediate reasoning behaviors of models are also discussed. Given an image, we restrict the perceptual field of a model, requiring it to actively zoom or shift its perceptual field based on reasoning to answer the question successfully. We conduct extensive evaluation over 30 models, including proprietary and open-source models, and observe that restricted perceptual fields play a significant role in enabling active perception. Results reveal a significant gap in the active perception capability of MLLMs, indicating that this area deserves more attention. We hope that ActiView could help develop methods for MLLMs to understand multimodal inputs in more natural and holistic ways.

Evaluating Theory of (an uncertain) Mind: Predicting the Uncertain Beliefs of Others from Conversational Cues

Anthony Sicilia, Malihe Alikhani

Typically, when evaluating Theory of Mind, we consider the beliefs of others to be binary: held or not held. But what if someone is unsure about their own beliefs? How can we quantify this uncertainty? We propose a new suite of tasks, challenging language models (LMs) to model the uncertainty of participants in a dialogue. We design these tasks around conversation forecasting, where the goal is to predict the probability of an unobserved conversation outcome. Uniquely, we view conversation agents themselves as forecasters, asking an LM to predict the uncertainty of an individual from their language use. We experiment with scaling methods, bagging, and demographic context for this regression task, conducting experiments on three dialogue corpora (social, negotiation, task-oriented) with eight LMs. While LMs can explain up to 7% variance in the uncertainty of others, we highlight the difficulty of the tasks and room for future work, especially in tasks that require explicit shifts in perspective.

Aligning VLM Assistants with Personalized Situated Cognition

Yongqi Li, Shen Zhou, Xiaohu Li, Xin Miao, Jintao Wen, Mayi Xu, Jianhao Chen, Birong Pan, Hankun Kang, Yuanyuan Zhu, Ming Zhong, Tieyun Qian

Vision-language models (VLMs) aligned with general human objectives, such as being harmless and hallucination-free, have become valuable assistants of humans in managing visual tasks. However, people with diversified backgrounds have different cognition even in the same situation. Consequently, they may have personalized expectations for VLM assistants. This highlights the urgent need to align VLM assistants with personalized situated cognition for real-world assistance. To study this problem, we first simplify it by characterizing individuals based on the sociological concept of Role-Set. Then, we propose to evaluate the individuals' actions to examine whether the personalized alignment is achieved. Further, we construct a benchmark named PCogAlignBench, which includes 18k instances and 20 individuals with different Role-Sets. Finally, we present a framework called PCogAlign, which constructs a cognition-aware and action-based reward model for personalized alignment. Experimental results and human evaluations demonstrate the reliability of the PCogAlignBench and the effectiveness of our proposed PCogAlign. We will open-source the constructed benchmark and code after being accepted.

HAIC: Improving Human Action Understanding and Generation with Better Captions for Multi-modal Large Language Models

Xiao Wang, Jingyun Hua, Weihong Lin, Yuanxing Zhang, Fuzheng Zhang, Jianlong Wu, Di ZHANG, Liqiang Nie

Recent Multi-modal Large Language Models (MLLMs) have made great progress in video understanding. However, their performance on videos involving human actions is still limited by the lack of high-quality data. To address this, we introduce a two-stage data annotation pipeline. First, we design strategies to accumulate videos featuring clear human actions from the Internet. Second, videos are annotated in a standardized caption format that uses human attributes to distinguish individuals and chronologically details their actions and interactions. Through this pipeline, we curate two datasets, namely HAICTrain and HAICBench. **HAICTrain** comprises 126K video-caption pairs generated by Gemini-Pro and verified for training purposes. Meanwhile, **HAICBench** includes 412 manually annotated video-caption pairs and 2,000 QA pairs, for a comprehensive evaluation of human action understanding. Experimental results demonstrate that training with HAICTrain not only significantly enhances human understanding abilities across 4 benchmarks, but can also improve text-to-video generation results. Both the HAICTrain and HAICBench will be made open-source to facilitate further research.

Disambiguating Reference in Visually Grounded Dialogues through Joint Modeling of Textual and Multimodal Semantic Structures

Shun Inadumi, Nobuhiro Ueda, Koichiro Yoshino

Multimodal reference resolution, including phrase grounding, aims to understand the semantic relations between mentions and real-world objects. Phrase grounding between images and their captions is a well-established task. In contrast, for real-world applications, it is essential to integrate textual and multimodal reference resolution to unravel the reference relations within dialogue, especially in handling ambiguities caused by pronouns and ellipses. This paper presents a framework that unifies textual and multimodal reference resolution by mapping mention embeddings to object embeddings and selecting mentions or objects based on their similarity. Our experiments show that learning textual reference resolution, such as coreference resolution and predicate-argument structure analysis, positively affects performance in multimodal reference resolution. In particular, our model with coreference resolution performs better in pronoun phrase grounding than representative models for this task, MDETR and GLIP. Our qualitative analysis demonstrates that incorporating textual reference relations strengthens the confidence scores between mentions, including pronouns and predicates, and objects, which can reduce the ambiguities that arise in visually grounded dialogues.

SPHERE: Unveiling Spatial Blind Spots in Vision-Language Models Through Hierarchical Evaluation

Wenyu Zhang, Wei En Ng, Lixin Ma, Yuwen Wang, Junqi Zhao, Allison Koenecke, Boyang Li, WANGLU

Current vision-language models may grasp basic spatial cues and simple directions (e.g. left, right, front, back), but struggle with the multi-dimensional spatial reasoning necessary for human-like understanding and real-world applications. To address this gap, we develop SPHERE (Spatial Perception and Hierarchical Evaluation of REasoning), a hierarchical evaluation framework supported by a new human-annotated dataset. SPHERE systematically probes models across increasing levels of complexity, from fundamental skills to multi-skill integration and high-level reasoning that combines spatial, visual, and logical understanding. Benchmark evaluation of state-of-the-art models reveals significant deficiencies, especially in reasoning about distance and proximity, understanding both egocentric and allocentric perspectives, and applying spatial logic in physical contexts. These findings expose critical blind spots in existing models and underscore the need for more advanced spatial reasoning techniques, driving the development of vision-language models that align more closely with human spatial cognition.

Enhancing Interpretable Image Classification Through LLM Agents and Conditional Concept Bottleneck Models

Yiwen Jiang, Deval Mehta, Wei Feng, Zongyuan Ge

Concept Bottleneck Models (CBMs) decompose image classification into a process governed by interpretable, human-readable concepts. Recent advances in CBMs have used Large Language Models (LLMs) to generate candidate concepts. However, a critical question remains: What is the optimal number of concepts to use? Current concept banks suffer from redundancy or insufficient coverage. To address this issue, we introduce a dynamic, agent-based approach that adjusts the concept bank in response to environmental feedback, optimizing the number of concepts for sufficiency yet concise coverage. Moreover, we propose Conditional Concept Bottleneck Models (CoCoBMs) to overcome the limitations in traditional CBMs' concept scoring mechanisms. It enhances the accuracy of assessing each concept's contribution to classification tasks and feature an editable matrix that allows LLMs to correct concept scores that conflict with their internal knowledge. Our evaluations across 6 datasets show that our method not only improves classification accuracy by 6% but also enhances interpretability assessments by 30%.

A Unified Agentic Framework for Evaluating Conditional Image Generation

Jifang Wang, Yangxue, Longyue Wang, Zhenran Xu, Yiyu Wang, Yaowei Wang, Weihua Luo, Kaifu Zhang, Baotian Hu, Min zhang

Conditional image generation has gained significant attention for its ability to personalize content. However, the field faces challenges in developing task-agnostic, reliable, and explainable evaluation metrics. This paper introduces CIGEval, a unified agentic framework for comprehensive evaluation of conditional image generation tasks. CIGEval utilizes large multimodal models (LMMs) as its core, integrating a multi-functional toolbox and establishing a fine-grained evaluation framework. Additionally, we synthesize evaluation trajectories for fine-tuning, empowering smaller LMMs to autonomously select appropriate tools and conduct nuanced analyses based on tool outputs. Experiments across seven prominent conditional image generation tasks demonstrate that CIGEval (GPT-4o version) achieves a high correlation of 0.4625 with human assessments, closely matching the inter-annotator correlation of 0.47. Notably, when implemented with 7B open-source LMMs using only 2.3K training trajectories, CIGEval surpasses the previous GPT-4o-based state-of-the-art method. These findings indicate that CIGEval holds great potential for automating evaluation of image generation tasks while maintaining human-level reliability.

Activation Steering Decoding: Mitigating Hallucination in Large Vision-Language Models through Bidirectional Hidden State Intervention

Jingran Su, Jingfan CHEN, Hongxin Li, Yuntao Chen, Li Qing, Zhaoxiang Zhang

Large Vision-Language Models (LVLMs) have demonstrated impressive capabilities in multimodal understanding, but they frequently suffer from hallucination - generating content inconsistent with visual inputs. In this work, we explore a novel perspective on hallucination mitigation by examining the intermediate activations of LVLMs during generation. Our investigation reveals that hallucinated content manifests as distinct, identifiable patterns in the model's hidden state space. Motivated by this finding, we propose Activation Steering Decoding (ASD), a training-free approach that mitigates hallucination through targeted intervention in the model's intermediate activations. ASD operates by first identifying directional patterns of hallucination in the activation space using a small calibration set, then employing a contrast decoding mechanism that computes the difference between positive and negative steering predictions. This approach effectively suppresses hallucination patterns while preserving the model's general capabilities. Extensive experiments demonstrate that our method significantly reduces hallucination across multiple benchmarks while maintaining performance on general visual understanding tasks. Notably, our approach requires no model re-training or architectural modifications, making it readily applicable to existing deployed models.

MapNav: A Novel Memory Representation via Annotated Semantic Maps for VLM-based Vision-and-Language Navigation

Lingfeng Zhang, Xiaoshuai Hao, Qinwen Xu, Qiang Zhang, Xinyao Zhang, Pengwei Wang, Jing Zhang, Zhongyuan Wang, Shanghang Zhang, Renjing Xu

Vision-language navigation (VLN) is a key task in Embodied AI, requiring agents to navigate diverse and unseen environments while following natural language instructions. Traditional approaches rely heavily on historical observations as spatio-temporal contexts for decision making, leading to significant storage and computational overhead. In this paper, we introduce MapNav, a novel end-to-end VLN model that leverages Annotated Semantic Map (ASM) to replace historical frames. Specifically, our approach constructs a top-down semantic map at the start of each episode and update it at each timestep, allowing for precise object mapping and structured navigation information. Then, we enhance this map with explicit textual labels for key regions, transforming abstract semantics into clear navigation cues and generate our ASM. MapNav agent using the constructed ASM as input, and use the powerful end-to-end capabilities of VLM to empower VLN. Extensive experiments demonstrate that MapNav achieves state-of-the-art (SOTA) performance in both simulated and real-world environments, validating the effectiveness of our method. We will release our ASM generation source code and dataset to ensure reproducibility, contributing valuable resources to the field. We believe that our proposed MapNav can be used as a new memory representation method in VLN, paving the way for future research in this field.

Exploring Compositional Generalization of Multimodal LLMs for Medical Imaging

Zhenyang Cai, Junying Chen, Rongsheng Wang, Weihong Wang, Yonglin Deng, Dingjie Song, Yize Chen, Zixu Zhang, Benyou Wang

Medical imaging provides essential visual insights for diagnosis, and multimodal large language models (MLLMs) are increasingly utilized for its analysis due to their strong generalization capabilities; however, the underlying factors driving this generalization remain unclear. Current research suggests that multi-task training outperforms single-task as different tasks can benefit each other, but they often overlook the internal relationships within these tasks. To analyze this phenomenon, we attempted to employ **compositional generalization** (CG), which refers to the models' ability to understand novel combinations by recombining learned elements, as a guiding framework. Since medical images can be precisely defined by **M**odality, **A**natomical area, and **T**ask, naturally providing an environment for exploring CG, we assembled 106 medical datasets to create **Med-MAT** for comprehensive experiments. The experiments confirmed that MLLMs can use CG to understand unseen medical images and identified CG as one of the main drivers of the generalization observed in multi-task training. Additionally, further studies demonstrated that CG effectively supports datasets with limited data and confirmed that MLLMs can achieve CG across classification and detection tasks, underscoring its broader generalization potential. Med-MAT is available at https://github.com/FreedomIntelligence/Med-MAT.

Do Multimodal Large Language Models Truly See What We Point At? Investigating Indexical, Iconic, and Symbolic Gesture Comprehension

Noriki Nishida, Koji Inoue, Hideki Nakayama, Mayumi Bono, Katsuya Takanashi

Understanding hand gestures is essential for human communication, yet it remains unclear how well multimodal large language models (MLLMs) comprehend them. In this paper, we examine MLLMs' ability to interpret indexical gestures, which require external referential grounding, in comparison to iconic gestures, which depict imagery, and symbolic gestures, which are conventionally defined. We hypothesize that MLLMs, lacking real-world referential understanding, will struggle significantly with indexical gestures. To test this, we manually annotated five gesture type labels to 925 gesture instances from the Miraikan SC Corpus and analyzed gesture descriptions generated by state-of-the-art MLLMs, including GPT-4o. Our findings reveal a consistent weakness across models in interpreting indexical gestures, suggesting that MLLMs rely heavily on linguistic priors or commonsense knowledge rather than grounding their interpretations in visual or contextual cues.

MM-Verify: Enhancing Multimodal Reasoning with Chain-of-Thought Verification

Linzhuang Sun, Hao Liang, Jingxuan Wei, Bihui Yu, Tianpeng Li, Fan Yang, Zenan Zhou, Wentao Zhang

According to the Test-Time Scaling, the integration of External Slow-Thinking with the Verify mechanism has been demonstrated to enhance multi-round reasoning in large language models (LLMs). However, in the multimodal (MM) domain, there is still a lack of a strong MM-Verifier. In this paper, we introduce MM-Verifier and MM-Reasoner to enhance multimodal reasoning through longer inference and more robust verification. First, we propose a two-step MM verification data synthesis method, which combines a simulation-based tree search with verification and uses rejection sampling to generate high-quality Chain-of-Thought (COT) data. This data is then used to fine-tune the verification model, MM-Verifier. Additionally, we present a more efficient method for synthesizing MMCOT data, bridging the gap between text-based and multimodal reasoning. The synthesized data is used to fine-tune MM-Reasoner. Our MM-Verifier outperforms all larger models on the MathCheck, MathVista, and MathVerse benchmarks. Moreover, MM-Reasoner demonstrates strong effectiveness and scalability, with performance improving as data size increases. Finally, our approach achieves strong performance when combining MM-Reasoner and MM-Verifier, reaching an accuracy of 65.3 on MathVista, surpassing GPT-4o (63.8) with 12 rollouts.

Can Vision-Language Models Evaluate Handwritten Math?

Oikantik Nath, Hanani Bathina, Mohammed Safi Ur Rahman Khan, Mitesh M Khapra

Recent advancements in Vision-Language Models (VLMs) have opened new possibilities in automatic grading of handwritten student responses, particularly in mathematics. However, a comprehensive study to test the ability of VLMs to evaluate and reason over handwritten content remains absent. To address this gap, we introduce FERMAT, a benchmark designed to assess VLMs' ability to detect, localize and correct errors in handwritten mathematical content. FERMAT spans four key error dimensions - computational, conceptual, notational, and presentation - and comprises over 2,200 handwritten math solutions derived from 609 manually curated problems from grades 7-12 with intentionally introduced perturbations. Using FERMAT we benchmark nine VLMs across three tasks: error detection, localization, and correction. Our results reveal significant shortcomings in current VLMs in reasoning over handwritten text, with Gemini-1.5-Pro achieving the highest error correction rate (77%). We also observed that some models struggle with processing handwritten content, as their accuracy improves when handwritten inputs are replaced with printed text or images. These findings highlight the limitations of current VLMs and reveal new avenues for improvement. We will release FERMAT and all the associated resources in the open-source to drive further research.

LLaVA Steering: Visual Instruction Tuning with 500x Fewer Parameters through Modality Linear Representation-Steering

Jinhe Bi, Yujun Wang, Haokun Chen, Xun Xiao, Artur Hecker, Volker Tresp, Yunpu Ma

Multimodal Large Language Models (MLLMs) enhance visual tasks by integrating visual representations into large language models (LLMs). The textual modality, inherited from LLMs, enables instruction following and in-context learning, while the visual modality boosts downstream task performance through rich semantic content, spatial information, and grounding capabilities. These modalities work synergistically across various visual tasks. Our research reveals a persistent imbalance between these modalities, with text often dominating output generation during visual instruction tuning, regardless of using full or parameter-efficient fine-tuning (PEFT). We found that re-balancing these modalities can significantly reduce trainable parameters, inspiring further optimization of visual instruction tuning. To this end, we introduce Modality Linear Representation-Steering (MoReS), which re-balances intrinsic modalities by steering visual representations through linear transformations in the visual subspace across each model layer. We validated our approach by developing LLaVA Steering, a suite of models using MoReS. Results show that LLaVA Steering requires, on average, 500 times fewer trainable parameters than LoRA while maintaining comparable performance across three visual benchmarks and eight visual question-answering tasks. Finally, we introduce the LLaVA Steering Factory, a platform that enables rapid customization of MLLMs with a component-based architecture, seamlessly integrating state-of-the-art models and evaluating intrinsic modality imbalance. This open-source project facilitates a deeper understanding of MLLMs within the research community.

Movie101v2: Improved Movie Narration Benchmark

Zihao Yue, Yepeng Zhang, Ziheng Wang, Qin Jin

Automatic movie narration aims to generate video-aligned plot descriptions to assist visually impaired audiences. Unlike standard video captioning, it involves not only describing key visual details but also inferring plots that unfold across multiple movie shots, presenting distinct and complex challenges. To advance this field, we introduce Movie101v2, a large-scale, bilingual dataset with enhanced data quality specifically designed for movie narration. Revisiting the task, we propose breaking down the ultimate goal of automatic movie narration into three progressive stages, offering a clear roadmap with corresponding evaluation metrics. Based on our new benchmark, we baseline a range of large vision-language models and conduct an in-depth analysis of the challenges in movie narration generation. Our findings highlight that achieving applicable movie narration generation is a fascinating goal that requires significant research.

Pixel-Level Reasoning Segmentation via Multi-turn Conversations

Dexian Cai, Xiaocui Yang, YongKang Liu, Daling Wang, Shi Feng, Yifei Zhang, Soujanya Poria

Existing visual perception systems focus on region-level segmentation in single-turn dialogues, relying on complex and explicit query instructions. Such systems cannot reason at the pixel level and comprehend dynamic user intent that changes over interaction. Our work tackles this issue by introducing a novel task, Pixel-level Reasoning Segmentation (Pixel-level RS) based on multi-turn conversations, tracking evolving user intent via multi-turn interactions for fine-grained segmentation. To establish a benchmark for this novel task, we build a Pixel-level ReasonIng Segmentation Dataset Based on Multi-Turn Conversations (PRIST), comprising 24k utterances from 8.3k multi-turn conversational scenarios with segmentation targets. Building on PRIST, we further propose MIRAS, a Multi-turn Interactive ReAsoning Segmentation framework, integrates pixel-level segmentation with robust multi-turn conversation understanding, generating pixel-grounded explanations aligned with user intent. The PRIST dataset and MIRSA framework fill the gap in pixel-level reasoning segmentation. Experimental results on the PRIST dataset demonstrate that our method outperforms current segmentation-specific baselines in terms of segmentation and LLM-based reasoning metrics. The code and data are available at: https://anonymous.4open.science/r/PixelRS/.

OmniAlign-V: Towards Enhanced Alignment of MLLMs with Human Preference

Xiangyu Zhao, Shengyuan Ding, Zicheng Zhang, Haian Huang, Maosongcao, Jiaqi Wang, Weiyun Wang, Xinyu Fang, Wenhai Wang, Guangtao Zhai, Hua Yang, Haodong Duan, Kai Chen

Recent advancements in open-source multi-modal large language models (MLLMs) have primarily focused on enhancing foundational capabilities, leaving a significant gap in human preference alignment. This paper introduces OmniAlign-V, a comprehensive dataset of 200K high-quality training samples featuring diverse images, complex questions, and varied response formats to improve MLLMs' alignment with human preferences. We also present MM-AlignBench, a human-annotated benchmark specifically designed to evaluate MLLMs' alignment with human values. Experimental results show that finetuning MLLMs with OmniAlign-V, using Supervised Fine-Tuning (SFT) or Direct Preference Optimization (DPO), significantly enhances human preference alignment while maintaining or enhancing performance on standard VQA benchmarks, preserving their fundamental capabilities.

V-Oracle: Making Progressive Reasoning in Deciphering Oracle Bones for You and Me

Runqi Qiao, Qiuna Tan, Guanting Dong, MinhuiWu, Jiapeng Wang, YiFan Zhang, Zhuoma GongQue, Chong Sun, Yida Xu, Yadong Xue, Ye Tian, Zhimin Bao, LAN YANG, Chen Li, Honggang Zhang

Oracle Bone Script (OBS) is a vital treasure of human civilization, rich in insights from ancient societies. However, the evolution of written language over millennia complicates its decipherment. In this paper, we propose V-Oracle, an innovative framework that utilizes Large Multi-modal Models (LMMs) for interpreting OBS. V-Oracle applies principles of pictographic character formation and frames the task as a visual question-answering (VQA) problem, establishing a multi-step reasoning chain. It proposes a multi-dimensional data augmentation for synthesizing high-quality OBS samples, and also implements a multi-phase oracle alignment tuning to improve LMMs' visual reasoning capabilities. Moreover, to bridge the evaluation gap in the OBS field, we further introduce Oracle-Bench, a comprehensive benchmark that emphasizes process-oriented assessment and incorporates both standard and out-of-distribution setups for realistic evaluation. Extensive experimental results can demonstrate the effectiveness of our method in providing quantitative analyses and superior deciphering capability.

Deep Temporal Reasoning in Video Language Models: A Cross-Linguistic Evaluation of Action Duration and Completion through Perfect Times

Olga Loginova, Sofía Ortega Loguinova

Human perception of events is intrinsically tied to distinguishing between completed (perfect and telic) and ongoing (durative) actions, a process mediated by both linguistic structure and visual cues. In this work, we introduce the Perfect Times dataset, a novel, quadrilingual (English, Italian, Russian, and Japanese) multiple-choice question-answering benchmark designed to assess video-language models (VLMs) on temporal reasoning. By pairing everyday activity videos with event completion labels and perfectivity-tailored distractors, our dataset probes whether models truly comprehend temporal dynamics or merely latch onto superficial markers. Experimental results indicate that state-of-the-art models, despite their success on text-based tasks, struggle to mirror human-like temporal and causal reasoning grounded in video. This study underscores the necessity of integrating deep multimodal cues to capture the nuances of action duration and completion within temporal and causal video dynamics, setting a new standard for evaluating and advancing temporal reasoning in VLMs.

SYNTHIA: Novel Concept Design with Affordance Composition

Hyeonjeong Ha, Xiaomeng Jin, Jeonghwan Kim, Jiateng Liu, Zhenhailong Wang, Khanh Duy Nguyen, Ansel Blume, Nanyun Peng, Kai-Wei Chang, Heng Ji

Text-to-image (T2I) models enable rapid concept design, making them widely used in AI-driven design. While recent studies focus on generating semantic and stylistic variations of given design concepts, --the integration of multiple affordances into a single coherent concept--remains largely overlooked. In this paper, we introduce SYNTHIA, a framework for generating novel, functionally coherent designs based on desired affordances. Our approach leverages a hierarchical concept ontology that decomposes concepts into parts and affordances, serving as a crucial building block for functionally coherent design. We also develop a curriculum learning scheme based on our ontology that contrastively fine-tunes T2I models to progressively learn affordance composition while maintaining visual novelty. To elaborate, we (i) gradually increase affordance distance, guiding models from basic concept-affordance association to complex affordance compositions that integrate parts of distinct affordances into a single, coherent form, and (ii) enforce visual novelty by employing contrastive objectives to push learned representations away from existing concepts. Experimental results show that SYNTHIA outperforms state-of-the-art T2I models, demonstrating absolute gains of 25.1% and 14.7% for novelty and functional coherence in human evaluation, respectively.

GUICourse: From General Vision Language Model to Versatile GUI Agent

Wentong Chen, Junbo Cui, Jinyi Hu, Yujia Qin, Junjie Fang, Yue Zhao, Chongyi Wang, Jun Liu, Guirong Chen, Yupeng Huo, Yuan Yao, Yankai Lin, Zhiyuan Liu, Maosong Sun

Utilizing Graphic User Interfaces (GUIs) for human-computer interaction is essential for accessing various digital tools. Recent advancements in Vision Language Models (VLMs) reveal significant potential for developing versatile agents that assist humans in navigating GUIs. However, current VLMs face challenges related to fundamental abilities, such as OCR and grounding, as well as a lack of knowledge about GUI elements functionalities and control methods. These limitations hinder their effectiveness as practical GUI agents. To address these challenges, we introduce GUICourse, a series of datasets for training visual-based GUI agents using general VLMs. First, we enhance the OCR and grounding capabilities of VLMs using the GUIEnv dataset. Next, we enrich the GUI knowledge of VLMs using the GUIAct and GUIChat datasets. Our experiments demonstrate that even a small-sized GUI agent (with 3.1 billion parameters) performs effectively on both single-step and multi-step GUI tasks. We further finetune our GUI agents on other GUI tasks with different action spaces (AITW and Mind2Web), and the results show that our agents are better than their baseline VLMs. Additionally, we analyze the impact of OCR and grounding capabilities through an ablation study, revealing a positive correlation with GUI navigation ability.

EffiVLM-Bench: A Comprehensive Benchmark for Evaluating Training-Free Acceleration in Large Visual-Languge Models

Zekun Wang, MingHua Ma, Zexin Wang, Rongchuan Mu, hongtao liu, liping shan, Ming Liu, Bing Qin

Large Vision-Language Models (LVLMs) have achieved remarkable success, yet their significant computational demands hinder practical deployment. While efforts to improve LVLM efficiency are growing, existing methods lack comprehensive evaluation across diverse backbones, benchmarks, and metrics. In this work, we systematically evaluate mainstream acceleration techniques for LVLMs, categorized into token and parameter compression. We introduce EffiVLM-BENCH, a unified framework for assessing not only absolute performance but also generalization and loyalty, while exploring Pareto-optimal trade-offs. Our extensive experiments and in-depth analyses offer insights into optimal strategies for accelerating LVLMs. We open-source code and recipes for EffiVLM-BENCH to foster future research.

CrafText Benchmark: Advancing Language Grounding in Complex Multimodal Open-Ended World

Zoya Volovikova, Gregory Gorbov, Petr Kuderov, Aleksandr Panov, Alexey Skrynnik

Following instructions in real-world conditions requires a capability to adapt to the world's volatility and entanglement: the environment is dynamic and unpredictable, instructions can be linguistically complex with diverse vocabulary, and the number of possible goals an agent may encounter is vast. Despite extensive research in this area, most studies are conducted in static environments with simple instructions and a limited vocabulary, making it difficult to assess agent performance in more diverse and challenging settings. To address this gap, we introduce CrafText, a benchmark for evaluating instruction following in a multimodal environment with diverse instructions and dynamic interactions. CrafText includes 3,924 instructions with 3,423 unique words, covering Localization, Conditional, Building, and Achievement tasks. Additionally, we propose an evaluation protocol that measures an agent’s ability to generalize to novel instruction formulations and dynamically evolving task configurations, providing a rigorous test of both linguistic understanding and adaptive decision-making.

T2I-FactualBench: Benchmarking the Factuality of Text-to-Image Models with Knowledge-Intensive Concepts

Ziwei Huang, Wanggui He, Quanyu Long, Yandi Wang, Haoyuan Li, Zhelun Yu, Fangxun Shu, Weilong Dai, Hao Jiang, Leilei Gan, Fei Wu

Most existing studies on evaluating text-to-image (T2I) models primarily focus on evaluating text-image alignment, image quality, and object composition capabilities, with comparatively fewer studies addressing the evaluation of the factuality of the synthesized images, particularly when the images involve knowledge-intensive concepts. In this work, we present T2I-FactualBench—the largest benchmark to date in terms of the number of concepts and prompts specifically designed to evaluate the factuality of knowledge-intensive concept generation. T2I-FactualBench consists of a three-tiered knowledge-intensive text-to-image generation framework, ranging from the basic memorization of individual knowledge concepts to the more complex composition of multiple knowledge concepts. We further introduce a multi-round visual question answering (VQA)-based evaluation framework to assesses the factuality of three-tiered knowledge-intensive text-to-image generation tasks. Experiments on T2I-FactualBench indicate that current state-of-the-art (SOTA) T2I models still leave significant room for improvement. We release our datasets and code at https://github.com/Safeoffellow/T2I-FactualBench.

Weaving Context Across Images: Improving Vision-Language Models through Focus-Centric Visual Chains

Juntian Zhang, Chuanqi Cheng, Yuhan Liu, Wei Liu, Jian Luan, Rui Yan

Vision-language models (VLMs) achieve remarkable success in single-image tasks. However, real-world scenarios often involve intricate multi-image inputs, leading to a notable performance decline as models struggle to disentangle critical information scattered across complex visual features. In this work, we propose Focus-Centric Visual Chain, a novel paradigm that enhances VLMs’ perception, comprehension, and reasoning abilities in multi-image scenarios. To facilitate this paradigm, we propose Focus-Centric Data Synthesis, a scalable bottom-up approach for synthesizing high-quality data with elaborate reasoning paths. Through this approach, We construct VISC-150K, a large-scale dataset with reasoning data in the form of Focus-Centric Visual Chain, specifically designed for multi-image tasks. Experimental results on seven multi-image benchmarks demonstrate that our method achieves average performance gains of 3.16% and 2.24% across two distinct model architectures, without compromising the general vision-language capabilities. Our study represents a significant step toward more robust and capable vision-language systems that can handle complex visual scenarios.

EditInspector: A Benchmark for Evaluation of Text-Guided Image Edits

Ron Yosef, Yonatan Bitton, Dani Lischinski, Moran Yanuka

Text-guided image editing, fueled by recent advancements in generative AI, is becoming increasingly widespread. This trend highlights the need for a comprehensive framework to verify text-guided edits and assess their quality. To address this need, we introduce EditInspector, a novel benchmark for evaluation of text-guided image edits, based on human annotations collected using an extensive template for edit verification. We leverage EditInspector to evaluate the performance of state-of-the-art (SoTA) vision and language models in assessing edits across various dimensions, including accuracy, artifact detection, visual quality, seamless integration with the image scene, adherence to common sense, and the ability to describe edit-induced changes. Our findings indicate that current models struggle to evaluate edits comprehensively and frequently hallucinate when describing the changes. To address these challenges, we propose two novel methods that outperform SoTA models in both artifact detection and difference caption generation.

Predicting Implicit Arguments in Procedural Video Instructions

Anil Batra, Laura Sevilla-Lara, Marcus Rohrbach, Frank Keller

Procedural texts help AI enhance reasoning about context and action sequences. Transforming these into Semantic Role Labeling (SRL) improves understanding of individual steps by identifying predicate-argument structure like {verb,what,where/with}. Procedural instructions are highly elliptic, for instance, (i) add cucumber to the bowl and (ii) add sliced tomatoes, the second step's where argument is inferred from the context, referring to where the cucumber was placed. Prior SRL benchmarks often miss implicit arguments, leading to incomplete understanding. To address this, we introduce Implicit-VidSRL, a dataset that necessitates inferring implicit and explicit arguments from contextual information in multimodal cooking procedures. Our proposed dataset benchmarks multimodal models' contextual reasoning, requiring entity tracking through visual changes in recipes. We study recent multimodal LLMs and reveal that they struggle to predict implicit arguments of what and where/with from multi-modal procedural data given the verb. Lastly, we propose iSRL-Qwen2-VL, which achieves a 17% relative improvement in F1-score for what-implicit and a 14.7% for where/with-implicit semantic roles over GPT-4o.

ViGiL3D: A Linguistically Diverse Dataset for 3D Visual Grounding

Austin Wang, ZeMing Gong, Angel X Chang

3D visual grounding (3DVG) involves localizing entities in a 3D scene referred to by natural language text. Such models are useful for embodied AI and scene retrieval applications, which involve searching for objects or patterns using natural language descriptions. While recent works have focused on LLM-based scaling of 3DVG datasets, these datasets do not capture the full range of potential prompts which could be specified in the English language. To ensure that we are scaling up and testing against a useful and representative set of prompts, we propose a framework for linguistically analyzing 3DVG prompts and introduce Visual Grounding with Diverse Language in 3D (ViGiL3D), a diagnostic dataset for evaluating visual grounding methods against a diverse set of language patterns. We evaluate existing open-vocabulary 3DVG methods to demonstrate that these methods are not yet proficient in understanding and identifying the targets of more challenging, out-of-distribution prompts, toward real-world applications.

Benchmarking and Improving Large Vision-Language Models for Fundamental Visual Graph Understanding and Reasoning

Yingjie Zhu, Xuefeng Bai, Kehai Chen, Yang Xiang, Jun Yu, Min Zhang

Large Vision-Language Models (LVLMs) have demonstrated remarkable performance across diverse tasks. Despite great success, recent studies show that LVLMs encounter substantial limitations when engaging with visual graphs. To study the reason behind these limitations, we propose VGCure, a comprehensive benchmark covering 22 tasks for examining the fundamental graph understanding and reasoning capacities of LVLMs. Extensive evaluations conducted on 14 LVLMs reveal that LVLMs are weak in basic graph understanding and reasoning tasks, particularly those concerning relational or structurally complex information. Based on this observation, we propose a structure-aware fine-tuning framework to enhance LVLMs with structure learning abilities through three self-supervised learning tasks. Experiments validate the effectiveness of our method in improving LVLMs' performance on fundamental and downstream graph learning tasks, as well as enhancing their robustness against complex visual graphs.

Activating Distributed Visual Region within LLMs for Efficient and Effective Vision-Language Training and Inference

Siyuan Wang, Dianyi Wang, Chengxing Zhou, Zejun Li, Zhihao Fan, Xuanjing Huang, zhongyu wei

Large Vision-Language Models (LVLMs) typically learn visual capacity through visual instruction tuning, involving updates to both a projector and their LLM backbones. Inspired by the concept of a visual region in the human brain, we investigate the existence of an analogous \textit{visual region} within LLMs that functions as a cognitive core, and explore the potential of efficient training of LVLMs via selective layers tuning. Using Bunny-Llama-3-8B-V for detailed analysis and other three LVLMs for validation across diverse visual and textual tasks, we find that selectively updating 25% of LLMs layers, when sparsely and uniformly distributed, can preserve nearly 99% of visual performance and maintain or improve textual task results, while effectively reducing training time. Based on this targeted training approach, we further propose a novel visual region-based pruning paradigm, removing non-critical layers outside the visual region, which can achieve minimal performance loss. This study offers an effective and efficient strategy for LVLM training and inference by activating a layer-wise visual region within LLMs, which proves consistently effective across different models.

REAL-MM-RAG: A Real-World Multi-Modal Retrieval Benchmark

navve wasserman, roi pony, Oshri Naparstek, Adi Raz Goldfarb, Eli Schwartz, Udi Barzelay, Leonid Karlinsky

Accurate multi-modal document retrieval is crucial for Retrieval-Augmented Generation (RAG), yet existing benchmarks do not fully capture real-world challenges with their current design. We introduce REAL-MM-RAG, an automatically generated benchmark designed to address four key properties essential for real-world retrieval: (i) multi-modal documents, (ii) enhanced difficulty, (iii) Realistic-RAG queries and (iv) accurate labeling. Additionally, we propose a multi-difficulty-level scheme based on query rephrasing to evaluate models' semantic understanding beyond keyword matching. Our benchmark reveals significant model weaknesses, particularly in handling table-heavy documents and robustness to query rephrasing. To mitigate these shortcomings, we curate a rephrased training set and introduce a new finance-focused, table-heavy dataset. Fine-tuning on these datasets enables models to achieve state-of-the-art retrieval performance on REAL-MM-RAG benchmark. Our work offers a better way to evaluate and improve retrieval in multi-modal RAG systems while also providing training data and models that address current limitations.

Session 2: NLP Applications (Monday 28 July, 11:00-12:30, Hall 4/5)

Literature Meets Data: A Synergistic Approach to Hypothesis Generation

Haokun Liu, Yangqiaoyu Zhou, Mingxuan Li, Chenfei Yuan, Chenhao Tan

AI holds promise for transforming scientific processes, including hypothesis generation. Prior work on hypothesis generation can be broadly categorized into theory-driven and data-driven approaches. While both have proven effective in generating novel and plausible hypotheses, it remains an open question whether they can complement each other. To address this, we develop the first method that combines literature-based insights with data to perform LLM-powered hypothesis generation. We apply our method on five different datasets and demonstrate that integrating literature and data outperforms other baselines (8.97% over few-shot, 15.75% over literature-based alone, and 3.37% over data-driven alone). Additionally, we conduct the first human evaluation to assess the utility of LLM-generated hypotheses in assisting human decision-making on two challenging tasks: deception detection and AI generated content detection. Our results show that human accuracy improves significantly by 7.44% and 14.19% on these tasks, respectively. These findings suggest that integrating literature-based and data-driven approaches provides a comprehensive and nuanced framework for hypothesis generation and could open new avenues for scientific inquiry.

LLM-Powered Test Case Generation for Detecting Bugs in Plausible Programs

Kaibo Liu, Zhenpeng Chen, Yiyang Liu, Jie Zhang, Mark Harman, Yudong Han, Yun Ma, Yihong Dong, Ge Li, Gang Huang

Detecting tricky bugs in plausible programs, those that pass existing test suites yet still contain bugs, remains a significant challenge in software testing. To address this problem, we propose TrickCatcher, an LLM-powered approach to generating test cases for uncovering bugs in plausible programs. TrickCatcher operates in three stages: First, it uses an LLM to generate program variants based on the program under test (PUT) and its specification. Second, it employs an LLM to construct an input generator from the specification for producing test inputs. Finally, these inputs are executed on both the PUT and its program variants to detect inconsistencies in their outputs. We evaluate TrickCatcher on two datasets, TrickyBugs and EvalPlus, which include 366 human-written and 151 AI-generated plausible programs with tricky bugs. TrickCatcher achieves recall, precision, and F1 scores that are 1.80×, 2.65×, and 1.66× those of the state-of-the-art baselines, respectively. Code and data used are available at https://github.com/RinCloud/TrickCatcher/.

A Survey on Foundation Language Models for Single-cell Biology

Fan Zhang, Hao Chen, Zhihong Zhu, Ziheng Zhang, Zhenxi Lin, Ziyue Qiao, Yefeng Zheng, Xian Wu

The recent advancements in language models have significantly catalyzed progress in computational biology. A growing body of research strives to construct unified foundation models for single-cell biology, with language models serving as the cornerstone. In this paper, we systematically review the developments in foundation language models designed specifically for single-cell biology. Our survey offers a thorough analysis of various incarnations of single-cell foundation language models, viewed through the lens of both pre-trained language models (PLMs) and large language models (LLMs). This includes an exploration of data tokenization strategies, pre-training/tuning paradigms, and downstream single-cell data analysis tasks. Additionally, we discuss the current challenges faced by these pioneering works and speculate on future research directions. Overall, this survey provides a comprehensive overview of the existing single-cell foundation language models, paving the way for future research endeavors.

Personality-Guided Code Generation Using Large Language Models

Yaoqi Guo, Zhenpeng Chen, Jie Zhang, Yang Liu, Yun Ma

Code generation, the automatic creation of source code from natural language descriptions, has garnered significant attention due to its potential to streamline software development. Inspired by research that links task-personality alignment with improved development outcomes, we conduct an empirical study on personality-guided code generation using large language models (LLMs). Specifically, we investigate how emulating personality traits appropriate to the coding tasks affects LLM performance. We extensively evaluate this approach using seven widely adopted LLMs across four representative datasets. Our results show that personality guidance significantly enhances code generation accuracy, with improved pass rates in 23 out of 28 LLM-dataset combinations. Notably, in 11 cases, the improvement exceeds 5%, and in 5 instances, it surpasses 10%, with the highest gain reaching 12.9%. Additionally, personality guidance can be easily integrated with other prompting strategies to further boost performance.

Re-identification of De-identified Documents with Autoregressive Infilling

Lucas Georges Gabriel Charpentier, Pierre Lison

Documents revealing sensitive information about individuals must typically be de-identified. This de-identification is often done by masking all mentions of personally identifiable information (PII), thereby making it more difficult to uncover the identity of the person(s) in question. To investigate the robustness of de-identification methods, we present a novel, RAG-inspired approach that attempts the reverse process of re-identification based on a database of documents representing background knowledge. Given a text in which personal identifiers have been masked, the re-identification proceeds in two steps. A retriever first selects from the background knowledge passages deemed relevant for the re-identification. Those passages are then provided to an infilling model which seeks to infer the original content of each text span. This process is repeated until all masked spans are replaced. We evaluate the re-identification on three datasets (Wikipedia biographies, court rulings and clinical notes). Results show that (1) as many as 80% of de-identified text spans can be successfully recovered and (2) the re-identification accuracy increases along with the level of background knowledge.

LEANCODE: Understanding Models Better for Code Simplification of Pre-trained Large Language Models

YAN WANG, Ling Ding, Tien N Nguyen, Shaohua Wang, Yanan Zheng

Large Language Models for code often entail significant computational complexity, which grows significantly with the length of the input code sequence. We propose LeanCode for code simplification to reduce training and prediction time, leveraging code contexts in utilizing attention scores to represent the tokens' importance. We advocate for the selective removal of tokens based on the average context-aware attention scores rather than average scores across all inputs. LeanCode uses the attention scores of `CLS' tokens within the encoder for classification tasks, such as code search. It also employs the encoder-decoder attention scores to determine token significance for sequence-to-sequence tasks like code summarization. Our evaluation shows LeanCode's superiority over the SOTAs DietCode and SlimCode, with improvements of 60\% and 16\% for code search, and 29\% and 27\% for code summarization, respectively.

Ask-Before-Detection: Identifying and Mitigating Conformity Bias in LLM-Powered Error Detector for Math Word Problem Solutions

Hang Li, Tianlong Xu, Kaiqi Yang, Yucheng Chu, Yanling Chen, Yichi Song, Qingsong Wen, Hui Liu

The rise of large language models (LLMs) offers new opportunities for automatic error detection in education, particularly for math word problems (MWPs). While prior studies demonstrate the promise of LLMs as error detectors, they overlook the presence of multiple valid solutions for a single MWP. Our preliminary analysis reveals a significant performance gap between conventional and alternative solutions in MWPs, a phenomenon we term conformity bias in this work. To mitigate this bias, we introduce the Ask-Before-Detect (AskBD) framework, which generates adaptive reference solutions using LLMs to enhance error detection. Experiments on 200 examples of GSM8K show that AskBD effectively mitigates bias and improves performance, especially when combined with reasoning-enhancing techniques like chain-of-thought prompting.

From Information to Insight: Leveraging LLMs for Open Aspect-Based Educational Summarization

Yang Zhong, Diane Litman

This paper addresses the challenge of aspect-based summarization in education by introducing Reflective ASPect-based summarization (ReflectASP), a novel dataset that summarizes student reflections on STEM lectures. Despite the promising performance of large language models in general summarization, their application to nuanced aspect-based summaries remains under-explored. ReflectASP eases the exploration of open-aspect-based summarization (OABS), overcoming the limitations of current datasets and comes with ample human annotations. We benchmarked different types of zero-shot summarization methods and proposed two refinement methods to improve summaries, supported by both automatic and human manual evaluations. Additionally, we analyzed suggestions and revisions made during the refinement process, offering a fine-grained study of the editing strategies employed by these methods. We make our models, dataset, and all human evaluation results available at https://github.com/cs329yangzhong/ReflectASP.

CompileAgent: Automated Real-World Repo-Level Compilation with Tool-Integrated LLM-based Agent System

Li Hu, Guoqiang Chen, Xiuwei Shang, Shaoyin Cheng, Benlong Wu, LiGangyang, Xu Zhu, Weiming Zhang, Nenghai Yu

With open-source projects growing in size and complexity, manual compilation becomes tedious and error-prone, highlighting the need for automation to improve efficiency and accuracy. However, the complexity of compilation instruction search and error resolution makes automatic compilation challenging. Inspired by the success of LLM-based agents in various fields, we propose CompileAgent, the first LLM-based agent framework dedicated to repo-level compilation. CompileAgent integrates five tools and a flow-based agent strategy, enabling interaction with software artifacts for compilation instruction search and error resolution. To measure the effectiveness of our method, we design a public repo-level benchmark CompileAgentBench, and we also design two baselines for comparison by combining two compilation-friendly schemes. The performance on this benchmark shows that our method significantly improves the compilation success rate, ranging from 10% to 71%. Meanwhile, we evaluate the performance of CompileAgent under different agent strategies and verify the effectiveness of the flow-based strategy. Additionally, we emphasize the scalability of CompileAgent, further expanding its application prospects. The complete code and data are available at https://github.com/Ch3nYe/AutoCompiler.

AndroidLab: Developing and Evaluating Android Agents in A Reproducible Environment

Yifan Xu, Xiao Liu, Xueqiao Sun, Siyi Cheng, Hao Yu, Hanyu Lai, Shudan Zhang, Dan Zhang, Jie Tang, Yuxiao Dong

Autonomous agents have become increasingly important for interacting with the real world. Android agents, in particular, have been a frequently-mentioned interaction method. However, existing studies for training and evaluating Android agents lack systematic research on both open-source and closed-source models. In this work, we propose AndroidLab as a systematic Android agent framework. It includes an operation environment with different modalities, action space, and a reproducible benchmark. It supports both large language models (LLMs) and multimodal models (LMMs) in the same action space. AndroidLab benchmark includes predefined Android virtual devices and 138 tasks across nine apps built on these devices. By using the AndroidLab environment, we develop an Android Instruction dataset and train six open-source LLMs and LMMs, lifting the average success rates from 4.59\% to 21.50\% for LLMs and from 1.93\% to 13.28\% for LMMs. AndroidLab is open-sourced and publicly available at https://github.com/THUDM/Android-Lab.

INVESTORBENCH: A Benchmark for Financial Decision-Making Tasks with LLM-based Agent

Haohang Li, Yupeng Cao, Yangyang Yu, Shashidhar Reddy Javaji, Zhiyang Deng, Yueru He, Yuechen Jiang, Zining Zhu, K.P. Subbalakshmi, Jimin Huang, Lingfei Qian, Xueqing Peng, Jordan W. Suchow, Qianqian Xie

Recent advancements have underscored the potential of large language model (LLM)-based agents in financial decision-making. Despite this progress, the field currently encounters two main challenges: (1) the lack of a comprehensive LLM agent framework adaptable to a variety of financial tasks, and (2) the absence of standardized benchmarks and consistent datasets for assessing agent performance. To tackle these issues, we introduce \textsc{InvestorBench}, the first benchmark specifically designed for evaluating LLM-based agents in diverse financial decision-making contexts. \textsc{InvestorBench} enhances the versatility of LLM-enabled agents by providing a comprehensive suite of tasks applicable to different financial products, including single equities like stocks and cryptocurrencies, and exchange-traded funds (ETFs). Additionally, we assess the reasoning and decision-making capabilities of our agent framework using thirteen different LLMs as backbone models, across various market environments and tasks. Furthermore, we have curated a diverse collection of open-source, datasets and developed a comprehensive suite of environments for financial decision-making. This establishes a highly accessible platform for evaluating financial agents' performance across various scenarios.

AndroidGen: Building an Android Language Agent under Data Scarcity

Hanyu Lai, Junjie Gao, Xiao Liu, Yifan Xu, Shudan Zhang, Yuxiao Dong, Jie Tang

Large language models have opened up a world of possibilities for various NLP tasks, sparking optimism for the future. Despite their potential, LLMs have yet to be widely used as agents on real mobile devices. The main challenge is the need for high-quality data sources. Time constraints and labor intensity often hinder human annotation. On the other hand, existing LLMs exhibit inadequate completion rates and need a robust data filtration strategy. Given these challenges, we develop a framework called AndroidGen to enhance the capabilities of LLM-based agents under data scarcity. In addition, we leverage AndroidGen to collect trajectories given human tasks and train open-source LLMs on these trajectories to develop an open-source mobile agent without manually labeled trajectories. We extensively evaluate AndroidGen with AndroidWorld, AitW, and various popular applications, demonstrating its improvements and revealing potential areas for future improvement. Code, model, and data are available at https://github.com/THUDM/AndroidGen.

LADM: Long-context Training Data Selection with Attention-based Dependency Measurement for LLMs

Jianghao Chen, Junhong Wu, Yangyifan Xu, Jiajun Zhang

Long-context modeling has drawn more and more attention in the area of Large Language Models (LLMs). Continual training with long-context data becomes the de-facto method to equip LLMs with the ability to process long inputs. However, it still remains an open challenge to measure the quality of long-context training data. To address this issue, we propose a Long-context data selection framework with Attention-based Dependency Measurement (LADM), which can efficiently identify high-quality long-context data from a large-scale, multi-domain pre-training corpus. LADM leverages the retrieval capabilities of the attention mechanism to capture contextual dependencies, ensuring a comprehensive quality measurement of long-context data. Experimental results show that our LADM framework significantly boosts the performance of LLMs on multiple long-context tasks with only 1B tokens for continual training.

Detection of Human and Machine-Authored Fake News in Urdu

Muhammad Zain Ali, Yuxia Wang, Bernhard Pfahringer, Tony C Smith

The rise of social media has amplified the spread of fake news, now further complicated by large language models (LLMs) like ChatGPT, which ease the generation of highly convincing, error-free misinformation, making it increasingly challenging for the public to discern truth from falsehood. Traditional fake news detection methods relying on linguistic cues have also become less effective. Moreover, current detectors primarily focus on binary classification and English texts, often overlooking the distinction between machine-generated true vs. fake news and the detection in low-resource languages. To this end, we updated the detection schema to include machine-generated news focusing on Urdu. We further propose a conjoint detection strategy to improve the accuracy and robustness. Experiments show its effectiveness across four datasets in various settings.

An Efficient and Precise Training Data Construction Framework for Process-supervised Reward Model in Mathematical Reasoning

Wei Sun, Qianlong Du, Fuwei Cui, Jiajun Zhang

Enhancing the mathematical reasoning capabilities of Large Language Models (LLMs) is of great scientific and practical significance. Researchers typically employ process-supervised reward models (PRMs) to guide the reasoning process, effectively improving the models' reasoning abilities. However, existing methods for constructing process supervision training data, such as manual annotation and per-step Monte Carlo estimation, are often costly or suffer from poor quality. To address these challenges, this paper introduces a framework called EpicPRM (Efficient, Precise, Cheap), which annotates each intermediate reasoning step based on its quantified contribution and uses an adaptive binary search algorithm to enhance both annotation precision and efficiency. Using this approach, we efficiently construct a high-quality process supervision training dataset named Epic50k, consisting of 50k annotated intermediate steps. Compared to other publicly available datasets, the PRM trained on Epic50k demonstrates significantly superior performance.

Game Development as Human-LLM Interaction

Jiale Hong, Hongqiu Wu, hai zhao

Game development is a highly specialized task that relies on a complex game engine powered by complex programming languages, preventing many gaming enthusiasts from handling it. This paper introduces the \textit{Chat Game Engine (ChatGE)} powered by LLM, which allows everyone to develop a custom game using natural language through Human-LLM interaction. To enable an LLM to function as a ChatGE, we instruct it to perform the following processes in each turn: (1) P_{script}: configure the game script segment based on the user's input; (2) P_{code}: generate the corresponding code snippet based on the game script segment; (3) P_{utter}: interact with the user, including guidance and feedback. We propose a data synthesis pipeline based on LLM to generate game script-code pairs and interactions from a few manually crafted seed data. We propose a three-stage training strategy following curriculum learning principles to transfer the dialogue-based LLM to our ChatGE smoothly. We construct a ChatGE for poker games as a case study and comprehensively evaluate it from two perspectives: interaction quality and code correctness.

Text is All You Need: LLM-enhanced Incremental Social Event Detection

Zitai Qiu, Congbo Ma, Jia Wu, Jian Yang

Social event detection (SED) is the task of identifying, categorizing, and tracking events from social data sources such as social media posts, news articles, and online discussions. Existing state-of-the-art (SOTA) SED models predominantly rely on graph neural networks (GNNs), which involve complex graph construction and time-consuming training processes, limiting their practicality in real-world scenarios. In this paper, we rethink the key challenge in SED: the informal and noisy nature of short texts on social media platforms, which impacts clustering accuracy. We propose a novel framework, LLM-enhanced Social Event Detection (LSED), which leverages the rich background knowledge of large language models (LLMs) to address this challenge. Specifically, LSED utilizes LLMs to formalize and disambiguate short texts by completing abbreviations and summarizing informal expressions. Furthermore, we introduce hyperbolic space embeddings, which are more suitable for natural language sentence representations, to enhance clustering performance. Extensive experiments on two challenging real-world datasets demonstrate that LSED outperforms existing SOTA models, achieving improvements in effectiveness, efficiency, and stability. Our work highlights the potential of LLMs in SED and provides a practical solution for real-world applications.

MorphMark: Flexible Adaptive Watermarking for Large Language Models

Zongqi Wang, Tianle Gu, Baoyuan Wu, Yujiu Yang

Watermarking by altering token sampling probabilities based on red-green list is a promising method for tracing the origin of text generated by large language models (LLMs). However, existing watermark methods often struggle with a fundamental dilemma: improving watermark effectiveness (the detectability of the watermark) often comes at the cost of reduced text quality. This trade-off limits their practical application. To address this challenge, we first formalize the problem within a multi-objective trade-off analysis framework. Within this framework, we identify a key factor that influences the dilemma. Unlike existing methods, where watermark strength is typically treated as a fixed hyperparameter, our theoretical insights lead to the development of MorphMark—a method that adaptively adjusts the watermark strength in response to changes in the identified factor, thereby achieving an effective resolution of the dilemma. In addition, MorphMark also prioritizes flexibility since it is an model-agnostic and model-free watermark method, thereby offering a practical solution for real-world deployment, particularly in light of the rapid evolution of AI models. Extensive experiments demonstrate that MorphMark achieves a superior resolution of the effectiveness-quality dilemma, while also offering greater flexibility and time and space efficiency.

OS-Genesis: Automating GUI Agent Trajectory Construction via Reverse Task Synthesis

Qiushi Sun, Kanzhi Cheng, Zichen Ding, Chuanyang Jin, Yian Wang, Fangzhi Xu, Zhenyu Wu, Chengyou Jia, Liheng Chen, Zhoumianze Liu, Ben Kao, Guohao Li, Junxian He, Yu Qiao, Zhiyong Wu

Graphical User Interface (GUI) agents powered by Vision-Language Models (VLMs) have demonstrated human-like computer control capability. Despite their utility in advancing digital automation, the development of such agents faces a critical bottleneck: collecting high-quality trajectory data for training. Common practices for collecting such data rely on human supervision or synthetic data generation through executing pre-defined tasks, which are either resource-intensive or unable to guarantee data quality. Further, these approaches exhibit significant gaps between the generated data and online environments, alongside limited data diversity. To address this issue, we introduce OS-Genesis, a novel GUI data synthesis pipeline that overcomes the challenges above. Unlike prior methods that rely on preset tasks, OS-Genesis reverse engineers the GUI trajectory construction process. Agents first perceive environments and perform step-level interactions, then retrospectively derive high-quality tasks to enable trajectory-level exploration. A trajectory reward model is then employed to ensure the quality of the generated trajectories. We demonstrate that training GUI agents with OS-Genesis significantly improves their performance on highly challenging online benchmarks. In-depth analysis further validates OS-Genesis's cost-effectiveness and its superior data quality and diversity compared to existing synthesis methods.

Enhancing Multimodal Continual Instruction Tuning with BranchLoRA

Duzhen Zhang, Yong Ren, Zhong-Zhi Li, Yahan Yu, Jiahua Dong, Chenxing Li, Zhilong Ji, Jinfeng Bai

Multimodal Continual Instruction Tuning (MCIT) aims to finetune Multimodal Large Language Models (MLLMs) to continually align with human intent across sequential tasks. Existing approaches often rely on the Mixture-of-Experts (MoE) LoRA framework to preserve previous instruction alignments. However, these methods are prone to Catastrophic Forgetting (CF), as they aggregate all LoRA blocks via simple summation, which compromises performance over time. In this paper, we identify a critical parameter inefficiency in the MoELoRA framework within the MCIT context. Based on this insight, we propose BranchLoRA, an asymmetric framework to enhance both efficiency and performance. To mitigate CF, we introduce a flexible tuning-freezing mechanism within BranchLoRA, enabling branches to specialize in intra-task knowledge while fostering inter-task collaboration. Moreover, we incrementally incorporate task-specific routers to ensure an optimal branch distribution over time, rather than favoring the most recent task. To streamline inference, we introduce a task selector that automatically routes test inputs to the appropriate router without requiring task identity. Extensive experiments on the latest MCIT benchmark demonstrate that BranchLoRA significantly outperforms MoELoRA and maintains its superiority across various MLLM sizes.

UAlign: Leveraging Uncertainty Estimations for Factuality Alignment on Large Language Models

Boyang XUE, Fei Mi, Qi Zhu, Hongru WANG, Rui Wang, Sheng Wang, Erxin Yu, Xuming Hu, Kam-Fai Wong

Despite demonstrating impressive capabilities, Large Language Models (LLMs) still often struggle to accurately express the factual knowledge they possess, especially in cases where the LLMs' knowledge boundaries are ambiguous. To improve LLMs' factual expressions, we propose the UAlign framework, which leverages Uncertainty estimations to represent knowledge boundaries, and then explicitly incorporates these representations as input features into prompts for LLMs to Align with factual knowledge. First, we prepare the dataset on knowledge question-answering (QA) samples by calculating two uncertainty estimations, including confidence score and semantic entropy, to represent the knowledge boundaries for LLMs. Subsequently, using the prepared dataset, we train a reward model that incorporates uncertainty estimations and then employ the Proximal Policy Optimization (PPO) algorithm for factuality alignment on LLMs. Experimental results indicate that, by integrating uncertainty representations in LLM alignment, the proposed UAlign can significantly enhance the LLMs' capacities to confidently answer known questions and refuse unknown questions on both in-domain and out-of-domain tasks, showing reliability improvements and good generalizability over various prompt- and training-based baselines.

ProvBench: A Benchmark of Legal Provision Recommendation for Contract Auto-Reviewing

Xiuxuan Shen, Zhongyuan Jiang, Junsan Zhang, Junxiao Han, Yao Wan, Chengjie Guo, Bingcheng Liu, Jie Wu, Renxiang Li, Philip S. Yu

Contract review is a critical process to protect the rights and interests of the parties involved. However, this process is time-consuming, labor-intensive, and costly, especially when a contract faces multiple rounds of review. To accelerate the contract review and promote the completion of transactions, this paper introduces a novel benchmark of legal provision recommendation and conflict detection for contract auto-reviewing (ProvBench), which aims to recommend the legal provisions related to contract clauses and detect possible legal conflicts. Specifically, we construct the first Legal Provision Recommendation Dataset: ProvData, which covers 8 common contract types. In addition, we conduct extensive experiments to evaluate ProvBench on various state-of-the-art models. Experimental results validate the feasibility of ProvBench and demonstrate the effectiveness of ProvData. Finally, we identify potential challenges in the ProvBench and advocate for further investigation.

LexTempus: Enhancing Temporal Generalizability of Legal Language Models Through Dynamic Mixture of Experts

Santosh T.Y.S.S, Tuan-Quang Vuong

The rapid evolution of legal concepts over time necessitates that legal language models adapt swiftly accounting for the temporal dynamics. However, prior works have largely neglected this crucial dimension, treating legal adaptation as a static problem rather than a continuous process. To address this gap, we pioneer LexTempus, a dynamic mixture of experts model that explicitly models the temporal evolution of legal language in a parameter-efficient online learning framework. LexTempus starts with a single lightweight adapter expert and dynamically expands by adding new experts as significant deviations in the data distribution are detected. This self-expansion strategy allows LexTempus to adapt to new information without forgetting past knowledge, thereby improving temporal generalization. We use a a non-parametric similarity-based router to merge relevant experts into a unified expert for each test instance, ensuring efficient inference without additional overhead. We validate the effectiveness of LexTempus on ECHR and EU case law datasets, demonstrating its superiority in both perplexity and open-ended text generation quality metrics.

Personalized Text Generation with Contrastive Activation Steering

Jinghao Zhang, Yuting Liu, Wenjie Wang, Qiang Liu, Shu Wu, Liang Wang, Tat-Seng Chua

Personalized text generation aims to infer users' writing style preferences from their historical texts and generate outputs that faithfully reflect these stylistic characteristics. Existing solutions primarily adopt two paradigms: retrieval-augmented generation (RAG) and parameter-efficient fine-tuning (PEFT). While these approaches have advanced the field, they suffer from two critical limitations: (1) the entanglement of content semantics and stylistic patterns in historical texts impedes accurate modeling of user-specific writing preferences; and (2) scalability challenges arising from both RAG's inference latency by retrieval operations and PEFT's parameter storage requirements for per user model. To overcome these limitations, we propose StyleVector, a training-free framework that disentangles and represents personalized writing style as a vector in LLM's activation space, enabling style-steered generation during inference without requiring costly retrieval or parameter storage. Comprehensive experiments demonstrate that our framework achieves a significant 8% relative improvement in personalized generation while reducing storage requirements by 1700 \times over PEFT method.

DRAG: Distilling RAG for SLMs from LLMs to Transfer Knowledge and Mitigate Hallucination via Evidence and Graph-based Distillation

Jennifer Chen, Aidar Myrzakhan, Yaxin Luo, Hassaan Muhammad Khan, Sondos Mahmoud Bsharat, Zhiqiang Shen

Retrieval-Augmented Generation (RAG) methods have proven highly effective for tasks requiring factual consistency and robust knowledge retrieval. However, large-scale RAG systems consume significant computational resources and are prone to generating "hallucinated" content from Humans. In this work, we introduce DRAG, a novel framework for distilling RAG knowledge from large-scale Language Models (LLMs) into small LMs (SLMs). Our approach leverages evidence- and knowledge graph–based distillation, ensuring that the distilled model retains critical factual knowledge while significantly reducing model size and computational cost. By aligning the smaller model's predictions with a structured knowledge graph and ranked evidence, DRAG effectively mitigates hallucinations and improves factual accuracy. We further present a case demonstrating how our framework mitigates user privacy risks and introduce a corresponding benchmark. Experimental evaluations on multiple benchmarks demonstrate that our method outperforms the prior competitive RAG methods like MiniRAG for SLMs by up to 27.7% using the same models, preserving high-level efficiency and reliability. With DRAG, we provide a practical and resource-efficient roadmap to deploying enhanced retrieval and generation capabilities in small-size LLMs. Code is available at https://github.com/VILA-Lab/DRAG.

Attacking Vision-Language Computer Agents via Pop-ups

Yanzhe Zhang, Tao Yu, Diyi Yang

Autonomous agents powered by large vision and language models (VLM) have demonstrated significant potential in completing daily computer tasks, such as browsing the web to book travel and operating desktop software, which requires agents to understand these interfaces. Despite such visual inputs becoming more integrated into agentic applications, what types of risks and attacks exist around them still remain unclear. In this work, we demonstrate that VLM agents can be easily attacked by a set of carefully designed adversarial pop-ups, which human users would typically recognize and ignore. This distraction leads agents to click these pop-ups instead of performing their tasks as usual. Integrating these pop-ups into existing agent testing environments like OSWorld and VisualWebArena leads to an attack success rate (the frequency of the agent clicking the pop-ups) of 86% on average and decreases the task success rate by 47%. Basic defense techniques, such as asking the agent to ignore pop-ups or including an advertisement notice, are ineffective against the attack. Code is available at [this link](https://github.com/SALT-NLP/PopupAttack).

LocAgent: Agentic Code Localization with Graph-Based Indexing

Zhaoling Chen, Xiangru Tang, Gangda Deng, Fang Wu, Jialong Wu, Zhiwei Jiang, Viktor Prasanna, Arman Cohan, Xingyao Wang

Code localization--identifying precisely where in a codebase changes need to be made--is a fundamental yet challenging task in software maintenance. Existing approaches struggle to efficiently navigate complex codebases when identifying relevant code snippets. The challenge lies in bridging natural language problem descriptions with the target code elements, often requiring reasoning across hierarchical structures and multiple dependencies. We introduce \textsc{LocAgent}, a framework that addresses code localization through a graph-guided agent. By parsing codebases into directed heterogeneous graphs, \textsc{LocAgent} creates a lightweight representation that captures code structures and their dependencies, enabling LLM agents to effectively search and locate relevant entities through powerful multi-hop reasoning. Experimental results on real-world benchmarks demonstrate that our approach significantly enhances accuracy in code localization. Notably, our method with the fine-tuned \texttt{Qwen-2.5-Coder-Instruct-32B} model achieves comparable results to SOTA proprietary models at greatly reduced cost (approximately 86\% reduction), reaching up to 92.7\% accuracy on file-level localization while improving downstream GitHub issue resolution success rates by 12\% for multiple attempts (Pass@10). Our code is available at .

Are LLMs effective psychological assessors? Leveraging adaptive RAG for interpretable mental health screening through psychometric practice

Federico Ravenda, Seyed Ali Bahrainian, Andrea Raballo, Antonietta Mira, Noriko Kando

In psychological practice, standardized questionnaires serve as essential tools for assessing mental health through structured, clinically-validated questions (i.e., items). While social media platforms offer rich data for mental health screening, computational approaches often bypass these established clinical assessment tools in favor of black-box classification. We propose a novel questionnaire-guided screening framework that bridges psychological practice and computational methods through adaptive Retrieval-Augmented Generation (aRAG). Our approach links unstructured social media content and standardized clinical assessments by retrieving relevant posts for each questionnaire item and using Large Language Models (LLMs) to complete validated psychological instruments. Our findings demonstrate two key advantages of questionnaire-guided screening: First, when completing the Beck Depression Inventory-II (BDI-II), our approach matches or outperforms state-of-the-art performance on Reddit-based benchmarks without requiring training data. Second, we show that guiding LLMs through standardized questionnaires yields superior results compared to directly prompting them for depression screening. Additionally, we show as a proof-of-concept how our questionnaire-based methodology successfully extends to self-harm screening.

Focus on What Matters: Enhancing Medical Vision-Language Models with Automatic Attention Alignment Tuning

Aofei Chang, Le Huang, Alex James Boyd, Parminder Bhatia, Taha Kass-Hout, Cao Xiao, Fenglong Ma

Medical Large Vision-Language Models (Med-LVLMs) often exhibit suboptimal attention distribution on visual inputs, leading to hallucinated or inaccurate outputs. Existing methods primarily rely on inference-time interventions, which are limited in attention adaptation or require additional supervision. To address this, we propose A^3Tune, a novel fine-tuning framework for Automatic Attention Alignment Tuning. A Tune leverages zero-shot weak labels from SAM, refines them into prompt-aware labels using BioMedCLIP, and then selectively modifies visually-critical attention heads to improve alignment while minimizing interference. Additionally, we introduce a A^3MoE module, enabling adaptive parameter selection for attention tuning across diverse prompts and images. Extensive experiments on medical VQA and report generation benchmarks show that A^3Tune outperforms state-of-the-art baselines, achieving enhanced attention distributions and performance in Med-LVLMs.

Incorporating Domain Knowledge into Materials Tokenization

Yerim Oh, Jun-Hyung Park, Junho Kim, SungHo Kim, SangKeun Lee

While language models are increasingly utilized in materials science, typical models rely on frequency-centric tokenization methods originally developed for natural language processing. However, these methods frequently produce excessive fragmentation and semantic loss, failing to maintain the structural and semantic integrity of material concepts. To address this issue, we propose MATTER, a novel tokenization approach that integrates material knowledge into tokenization. Based on MatDetector trained on our materials knowledge base and re-ranking method prioritizing material terms in token merging, MATTER maintains the structural integrity of identified materials concepts and prevents fragmentation during tokenization, ensuring their semantic meaning remains intact. The experimental results demonstrate that MATTER outperforms existing tokenization methods, achieving an average performance gain of 4% and 2% in the generation and classification tasks, respectively. These results underscore the importance of domain knowledge for tokenization strategies in scientific text processing.

On the Risk of Evidence Pollution for Malicious Social Text Detection in the Era of LLMs

Herun Wan, Minnan Luo, Zhixiong Su, Guang Dai, Xiang Zhao

Evidence-enhanced detectors present remarkable abilities in identifying malicious social text. However, the rise of large language models (LLMs) brings potential risks of evidence pollution to confuse detectors. This paper explores potential manipulation scenarios including basic pollution, and rephrasing or generating evidence by LLMs. To mitigate the negative impact, we propose three defense strategies from the data and model sides, including machine-generated text detection, a mixture of experts, and parameter updating. Extensive experiments on four malicious social text detection tasks with ten datasets illustrate that evidence pollution significantly compromises detectors, where the generating strategy causes up to a 14.4% performance drop. Meanwhile, the defense strategies could mitigate evidence pollution, but they faced limitations for practical employment. Further analysis illustrates that polluted evidence (i) is of high quality, evaluated by metrics and humans; (ii) would compromise the model calibration, increasing expected calibration error up to 21.6%; and (iii) could be integrated to amplify the negative impact, especially for encoder-based LMs, where the accuracy drops by 21.8%.

S-RAG: A Novel Audit Framework for Detecting Unauthorized Use of Personal Data in RAG Systems

Zhirui Zeng, Jiamou Liu, Meng-Fen Chiang, Jialing He, Zijian Zhang

Retrieval-Augmented Generation (RAG) systems combine external data retrieval with text generation and have become essential in applications requiring accurate and context-specific responses. However, their reliance on external data raises critical concerns about unauthorized collection and usage of personal information. To ensure compliance with data protection regulations like GDPR and detect improper use of data, we propose the Shadow RAG Auditing Data Provenance (S-RAG) framework. S-RAG enables users to determine whether their textual data has been utilized in RAG systems, even in black-box settings with no prior system knowledge. It is effective across open-source and closed-source RAG systems and resilient to defense strategies. Experiments demonstrate that S-RAG achieves an improvement in Accuracy by 19.9% (compared to the best baseline), while maintaining strong performance under adversarial defenses. Furthermore, we analyze how the auditor's knowledge of the target system affects performance, offering practical insights for privacy-preserving AI systems. Our code is open-sourced online.

UniLR: Unleashing the Power of LLMs on Multiple Legal Tasks with a Unified Legal Retriever

Ang Li, Yiquan Wu, Yifei Liu, Ming Cai, Lizhi Qing, Shihang Wang, Yangyang Kang, Chengyuan Liu, Fei Wu, Kun Kuang

Despite the impressive capabilities of LLMs, they often generate content with factual inaccuracies in LegalAI, which may lead to serious legal consequences. Retrieval-Augmented Generation (RAG), a promising approach, can conveniently integrate specialized knowledge into LLMs. In practice, there are diverse legal knowledge retrieval demands (e.g. law articles and similar cases). However, existing retrieval methods are either designed for general domains, struggling with legal knowledge, or tailored for specific legal tasks, unable to handle diverse legal knowledge types. Therefore, we propose a novel **Uni**fied **L**egal **R**etriever (UniLR) capable of performing multiple legal retrieval tasks for LLMs. Specifically, we introduce attention supervision to guide the retriever in focusing on key elements during knowledge encoding. Next, we design a graph-based method to integrate meta information through a heterogeneous graph, further enriching the knowledge representation. These two components work together to enable UniLR to capture the essence of knowledge hidden beneath formats. Extensive experiments on multiple datasets of common legal tasks demonstrate that UniLR achieves the best retrieval performance and can significantly enhance the performance of LLM.

SURVEYFORGE : On the Outline Heuristics, Memory-Driven Generation, and Multi-dimensional Evaluation for Automated Survey Writing

Xiangchao Yan, Shiyang Feng, Jiakang Yuan, Renqiu Xia, Bin Wang, LEI BAI, Bo Zhang

Survey paper plays a crucial role in scientific research, especially given the rapid growth of research publications. Recently, researchers have begun using LLMs to automate survey generation for better efficiency. However, the quality gap between LLM-generated surveys and those written by human remains significant, particularly in terms of outline quality and citation accuracy. To close these gaps, we introduce SURVEYFORGE, which first generates the outline by analyzing the logical structure of human-written outlines and referring to the retrieved domain-related articles.Subsequently, leveraging high-quality papers retrieved from memory by our scholar navigation agent, SURVEYFORGE can automatically generate and refine the content of the generated article. Moreover, to achieve a comprehensive evaluation, we construct SurveyBench, which includes 100 human-written survey papers for win-rate comparison and assesses AI-generated survey papers across three dimensions: reference, outline, and content quality. Experiments demonstrate that SURVEYFORGEcan outperform previous works such as AutoSurvey.

Interactive Evolution: A Neural-Symbolic Self-Training Framework For Large Language Models

Fangzhi Xu, Qiushi Sun, Kanzhi Cheng, Jun Liu, Yu Qiao, Zhiyong Wu

One of the primary driving forces contributing to the superior performance of Large Language Models (LLMs) is the extensive availability of human-annotated natural language data, which is used for alignment fine-tuning. This inspired researchers to investigate self-training methods to mitigate the extensive reliance on human annotations. However, the current success of self-training has been primarily observed in natural language scenarios, rather than in the increasingly important neural-symbolic scenarios. To this end, we propose an environment-guided neural-symbolic self-training framework named ENVISIONS. It aims to overcome two main challenges: (1) the scarcity of symbolic data, and (2) the limited proficiency of LLMs in processing symbolic language. Extensive evaluations conducted on three distinct domains demonstrate the effectiveness of our approach. Additionally, we have conducted a comprehensive analysis to uncover the factors contributing to ENVISIONS's success, thereby offering valuable insights for future research in this area.

DREsS: Dataset for Rubric-based Essay Scoring on EFL Writing

Haneul Yoo, Jieun Han, So-Yeon Ahn, Alice Oh

Automated essay scoring (AES) is a useful tool in English as a Foreign Language (EFL) writing education, offering real-time essay scores for students and instructors. However, previous AES models were trained on essays and scores irrelevant to the practical scenarios of EFL writing education and usually provided a single holistic score due to the lack of appropriate datasets. In this paper, we release DREsS, a large-scale, standard dataset for rubric-based automated essay scoring with 48.9K samples in total. DREsS comprises three sub-datasets: DREsS_New, DREsS_Std., and DREsS_CASE. We collect DREsS_New, a real-classroom dataset with 2.3K essays authored by EFL undergraduate students and scored by English education experts. We also standardize existing rubric-based essay scoring datasets as DREsS_Std. We suggest CASE, a corruption-based augmentation strategy for essays, which generates 40.1K synthetic samples of DREsS_CASE and improves the baseline results by 45.44%. DREsS will enable further research to provide a more accurate and practical AES system for EFL writing education.

Self-supervised Quantized Representation for Seamlessly Integrating Knowledge Graphs with Large Language Models

Qika Lin, Tianzhe Zhao, Kai He, Zhen Peng, Fangzhi Xu, Ling Huang, Jingying Ma, Mengling Feng

Due to the presence of the natural gap between Knowledge Graph (KG) structures and the natural language, the effective integration of holistic structural information of KGs with Large Language Models (LLMs) has emerged as a significant question. To this end, we propose a two-stage framework to learn and apply quantized codes for each entity, aiming for the seamless integration of KGs with LLMs. Firstly, a self-supervised quantized representation (SSQR) method is proposed to compress both KG structural and semantic knowledge into discrete codes (i.e., tokens) that align the format of language sentences. We further design KG instruction-following data by viewing these learned codes as features to directly input to LLMs, thereby achieving seamless integration. The experiment results demonstrate that SSQR outperforms existing unsupervised quantized methods, producing more distinguishable codes. Moreover, the fine-tuned LLaMA2 and LLaMA3.1 also have superior performance on KG link prediction and triple classification tasks, utilizing only 16 tokens per entity instead of thousands in conventional prompting methods.

Enhancing Large Language Model’s Capabilities in Open Domains via Autonomous Tool Integration from GitHub

Bohan Lyu, Xin Cong, Heyang Yu, Pan Yang, Cheng Qian, Zihe Wang, Yujia Qin, Yining Ye, Yaxi Lu, Chen Qian, Zhong Zhang, Yukun Yan, Yankai Lin, Zhiyuan Liu, Maosong Sun

Large Language Models (LLMs) excel in traditional natural language processing tasks but struggle with problems that require complex domain-specific calculations or simulations. While equipping LLMs with external tools to build LLM-based agents can enhance their capabilities, existing approaches lack the flexibility to address diverse and ever-evolving user queries in open domains. Currently, there is also no existing dataset that evaluates LLMs on open-domain knowledge that requires tools to solve. To this end, we introduce OpenAct benchmark to evaluate the open-domain task-solving capability, which is built on human expert consultation and repositories in GitHub. It comprises 339 questions spanning 7 diverse domains that need to be solved with domain-specific methods. In our experiments, even state-of-the-art LLMs and LLM-based agents demonstrate unsatisfactory success rates, underscoring the need for a novel approach. Furthermore, we present OpenAgent, a novel LLM-based agent system that can tackle evolving queries in open domains through autonomously integrating specialized tools from GitHub. OpenAgent employs 1) a hierarchical framework where specialized agents handle specific tasks and can assign tasks to inferior agents, 2) a bi-level experience learning mechanism to learn from both humans' and its own experiences to tackle tool flaws. Experiments demonstrate its superior effectiveness and efficiency, which significantly outperforms baselines. Our data and code are open-source at https://github.com/OpenBMB/OpenAct.

SceneGenAgent: Precise Industrial Scene Generation with Coding Agent

Xiao Xia, Dan Zhang, Zibo Liao, Zhenyu Hou, Tianrui Sun, Jing Li, Ling Fu, Yuxiao Dong

The modeling of industrial scenes is essential for simulations in industrial manufacturing. While large language models (LLMs) have shown significant progress in generating general 3D scenes from textual descriptions, generating industrial scenes with LLMs poses a unique challenge due to their demand for precise measurements and positioning, requiring complex planning over spatial arrangement. To address this challenge, we introduce SceneGenAgent, an LLM-based agent for generating industrial scenes through C# code. SceneGenAgent ensures precise layout planning through a structured and calculable format, layout verification, and iterative refinement to meet the quantitative requirements of industrial scenarios. Experiment results demonstrate that LLMs powered by SceneGenAgent exceed their original performance, reaching up to 81.0% success rate in real-world industrial scene generation tasks and effectively meeting most scene generation requirements. To further enhance accessibility, we construct SceneInstruct, a dataset designed for fine-tuning open-source LLMs to integrate into SceneGenAgent. Experiments show that fine-tuning open-source LLMs on SceneInstruct yields significant performance improvements, with Llama3.1-70B approaching the capabilities of GPT-4o. Our code and dataset are available at https://github.com/THUDM/SceneGenAgent.

nvAgent: Automated Data Visualization from Natural Language via Collaborative Agent Workflow

Geliang Ouyang, Jingyao Chen, Zhihe Nie, Yi Gui, Yao Wan, Hongyu Zhang, Dongping Chen

Natural Language to Visualization (NL2Vis) seeks to convert natural-language descriptions into visual representations of given tables, empowering users to derive insights from large-scale data. Recent advancements in *Large Language Models* (LLMs) show promise in automating code generation to transform tabular data into accessible visualizations. However, they often struggle with complex queries that require reasoning across multiple tables. To address this limitation, we propose a collaborative agent workflow, termed **nvAgent**, for NL2Vis. Specifically, **nvAgent** comprises three agents: a processor agent for database processing and context filtering, a composer agent for planning visualization generation, and a validator agent for code translation and output verification. Comprehensive evaluations on the new VisEval benchmark demonstrate that **nvAgent** consistently surpasses state-of-the-art baselines, achieving a 7.88% improvement in single-table and a 9.23% improvement in multi-table scenarios. Qualitative analyses further highlight that **nvAgent** maintains nearly a 20% performance margin over previous models, underscoring its capacity to produce high-quality visual representations from complex, heterogeneous data sources. All datasets and source code are available at: https://github.com/geliang0114/nvAgent.

MaCP: Minimal yet Mighty Adaptation via Hierarchical Cosine Projection

Yixian Shen, Qi Bi, JIA-HONG HUANG, Hongyi Zhu, Andy D. Pimentel, Anuj Pathania

We present a new adaptation method MaCP, Minimal yet Mighty adaptive Cosine Projection, that achieves exceptional performance while requiring minimal parameters and memory for fine-tuning large foundation models. Its general idea is to exploit the superior energy compaction and decorrelation properties of cosine projection to improve both model efficiency and accuracy. Specifically, it projects the weight change from the low-rank adaptation into the discrete cosine space. Then, the weight change is partitioned over different levels of the discrete cosine spectrum, and each partition's most critical frequency components are selected. Extensive experiments demonstrate the effectiveness of MaCP across a wide range of single-modality tasks, including natural language understanding, natural language generation, text summarization, as well as multi-modality tasks such as image classification and video understanding. MaCP consistently delivers superior accuracy, significantly reduced computational complexity, and lower memory requirements compared to existing alternatives.

WET: Overcoming Paraphrasing Vulnerabilities in Embeddings-as-a-Service with Linear Transformation Watermarks

Anudeex Shetty, Qiongkai Xu, Jey Han Lau

Embeddings-as-a-Service (EaaS) is a service offered by large language model (LLM) developers to supply embeddings generated by LLMs. Previous research suggests that EaaS is prone to imitation attacks—attacks that clone the underlying EaaS model by training another model on the queried embeddings. As a result, EaaS watermarks are introduced to protect the intellectual property of EaaS providers. In this paper, we first show that existing EaaS watermarks can be removed by paraphrasing when attackers clone the model. Subsequently, we propose a novel watermarking technique that involves linearly transforming the embeddings, and show that it is empirically and theoretically robust against paraphrasing.

Generating Plausible Distractors for Multiple-Choice Questions via Student Choice Prediction

Yooseop Lee, Suin Kim, Yohan Jo

In designing multiple-choice questions (MCQs) in education, creating plausible distractors is crucial for identifying students' misconceptions and gaps in knowledge and accurately assessing their understanding. However, prior studies on distractor generation have not paid sufficient attention to enhancing the difficulty of distractors, resulting in reduced effectiveness of MCQs. This study presents a pipeline for training a model to generate distractors that are more likely to be selected by students. First, we train a pairwise ranker to reason about students' misconceptions and assess the relative plausibility of two distractors. Using this model, we create a dataset of pairwise distractor ranks and then train a distractor generator via Direct Preference Optimization (DPO) to generate more plausible distractors. Experiments on computer science subjects (Python, DB, MLDL) demonstrate that our pairwise ranker effectively identifies students' potential misunderstandings and achieves ranking accuracy comparable to human experts. Furthermore, our distractor generator outperforms several baselines in generating plausible distractors and produces questions with a higher item discrimination index (DI).

DEEPER Insight into Your User: Directed Persona Refinement for Dynamic Persona Modeling

Aili Chen, Chengyu Du, Jiangjie Chen, Jinghan Xu, Yikai Zhang, Siyu Yuan, Zulong Chen, Liangyue Li, Yanghua Xiao

To advance personalized applications such as recommendation systems and user behavior prediction, recent research increasingly adopts large language models (LLMs) for human-readable persona modeling. In dynamic real-world scenarios, effective persona modeling necessitates leveraging streaming behavior data to continually optimize user personas.However, existing methods—whether regenerating personas or incrementally extending them with new behaviors—often fail to achieve sustained improvements in persona quality or future behavior prediction accuracy. To address this, we propose DEEPER, a novel approach for dynamic persona modeling that enables continual persona optimization. Specifically, we enhance the model's direction-search capability through an iterative reinforcement learning framework, allowing it to automatically identify effective update directions and optimize personas using discrepancies between user behaviors and model predictions.Extensive experiments on dynamic persona modeling involving 4,800 users across 10 domains highlight \method's superior persona optimization capabilities, delivering an impressive 32.2% average reduction in user behavior prediction error over four update rounds—outperforming the best baseline by a remarkable 22.92%.

Asclepius: A Spectrum Evaluation Benchmark for Medical Multi-Modal Large Language Models

Jie Liu, Wenxuan Wang, SU Yihang, Jingyuan Huang, Yudi Zhang, Cheng-Yi Li, Wenting Chen, Xiaohan Xing, Kao-Jung Chang, Linlin Shen, Michael R. Lyu

The significant breakthroughs of Medical Multi-Modal Large Language Models (Med-MLLMs) renovate modern healthcare with robust information synthesis and medical decision support. However, these models are often evaluated on benchmarks that are unsuitable for the Med-MLLMs due to the intricate nature of the real-world diagnostic frameworks, which encompass diverse medical specialties and involve complex clinical decisions. Thus, a clinically representative benchmark is highly desirable for credible Med-MLLMs evaluation. To this end, we introduce Asclepius, a novel Med-MLLM benchmark that comprehensively assesses Med-MLLMs in terms of: distinct medical specialties (cardiovascular, gastroenterology, etc.) and different diagnostic capacities (perception, disease analysis, etc.). Grounded in 3 proposed core principles, Asclepius ensures a comprehensive evaluation by encompassing 15 medical specialties, stratifying into 3 main categories and 8 sub-categories of clinical tasks, and exempting overlap with the existing VQA dataset. We further provide an in-depth analysis of 6 Med-MLLMs and compare them with 3 human specialists, providing insights into their competencies and limitations in various medical contexts. Our work not only advances the understanding of Med-MLLMs' capabilities but also sets a precedent for future evaluations and the safe deployment of these models in clinical environments.

Knowledge Graph Retrieval-Augmented Generation for LLM-based Recommendation

Shijie Wang, Wenqi Fan, Yue Feng, LIN SHANRU, Xinyu Ma, Shuaiqiang Wang, Dawei Yin

Recommender systems have become increasingly vital in our daily lives, helping to alleviate the problem of information overload across various user-oriented online services. The emergence of Large Language Models (LLMs) has yielded remarkable achievements, demonstrating their potential for the development of next-generation recommender systems. Despite these advancements, LLM-based recommender systems face inherent limitations stemming from their LLM backbones, particularly issues of hallucinations and the lack of up-to-date and domain-specific knowledge. Recently, Retrieval-Augmented Generation (RAG) has garnered significant attention for addressing these limitations by leveraging external knowledge sources to enhance the understanding and generation of LLMs. However, vanilla RAG methods often introduce noise and neglect structural relationships in knowledge, limiting their effectiveness in LLM-based recommendations. To address these limitations, we propose to retrieve high-quality and up-to-date structure information from the knowledge graph (KG) to augment recommendations. Specifically, our approach develops a retrieval-augmented framework, termed K-RagRec, that facilitates the recommendation generation process by incorporating structure information from the external KG. Extensive experiments have been conducted to demonstrate the effectiveness of our proposed method.

Opt-Out: Investigating Entity-Level Unlearning for Large Language Models via Optimal Transport

Minseok Choi, Daniel Rim, Dohyun Lee, Jaegul Choo

Instruction-following large language models (LLMs), such as ChatGPT, have become widely popular among everyday users. However, these models inadvertently disclose private, sensitive information to their users, underscoring the need for machine unlearning techniques to remove selective information from the models. While prior work has focused on forgetting small, random subsets of training data at the instance-level, we argue that real-world scenarios often require the removal of an entire user data, which may require a more careful maneuver. In this study, we explore entity-level unlearning, which aims to erase all knowledge related to a target entity while preserving the remaining model capabilities. To address this, we introduce Opt-Out, an optimal transport-based unlearning method that utilizes the Wasserstein distance from the model's initial parameters to achieve more effective and fine-grained unlearning. We also present the first Entity-Level Unlearning Dataset (ELUDe) designed to evaluate entity-level unlearning. Our empirical results demonstrate that Opt-Out surpasses existing methods, establishing a new standard for secure and adaptable LLMs that can accommodate user data removal requests without the need for full retraining.

DISC: Plug-and-Play Decoding Intervention with Similarity of Characters for Chinese Spelling Check

Ziheng Qiao, Houquan Zhou, Yumeng Liu, Zhenghua Li, Min Zhang, Bo Zhang, Chen Li, Ji Zhang, Fei Huang

One key characteristic of the Chinese spelling check (CSC) task is that incorrect characters are usually similar to the correct ones in either phonetics or glyph. To accommodate this, previous works usually leverage confusion sets, which suffer from two problems, i.e., difficulty in determining which character pairs to include and lack of probabilities to distinguish items in the set. In this paper, we propose a light-weight plug-and-play DISC (i.e., decoding intervention with similarity of characters) module for CSC models. DISC measures phonetic and glyph similarities between characters and incorporates this similarity information only during the inference phase. This method can be easily integrated into various existing CSC models, such as ReaLiSe, SCOPE, and ReLM, without additional training costs. Experiments on three CSC benchmarks demonstrate that our proposed method significantly improves model performance, approaching and even surpassing the current state-of-the-art models.

Basic Reading Distillation

Zhi Zhou, Sirui Miao, Xiangyu Duan, Hao Yang, Min Zhang

Large language models (LLMs) have demonstrated remarkable abilities in various natural language processing areas, but they demand high computation resources which limits their deployment in real-world. Distillation is one technique to solve this problem through either knowledge distillation or task distillation. Both distillation approaches train small models to imitate specific features of LLMs, but they all neglect basic reading education for small models on generic texts that are \emph{unrelated} to downstream tasks. In this paper, we propose basic reading distillation (BRD) which educates a small model to imitate LLMs basic reading behaviors, such as named entity recognition, question raising and answering, on each sentence. After such basic education, we apply the small model on various tasks including language inference benchmarks and BIG-bench tasks. It shows that the small model can outperform or perform comparable to over 20x bigger LLMs. Analysis reveals that BRD effectively influences the probability distribution of the small model, and has orthogonality to either knowledge distillation or task distillation.

GPT-4 as a Homework Tutor Can Improve Student Engagement and Learning Outcomes

Alessandro Vanzo, Sankalan Pal Chowdhury, Mrinmaya Sachan

This work contributes to the scarce empirical literature on LLM-based interactive homework in real-world educational settings and offers a practical, scalable solution to improve homework in schools. Homework is an important part of education in schools across the world, but to maximize benefit, it must be accompanied by feedback and follow-up questions. We developed a prompting strategy that enables GPT-4 to conduct interactive homework sessions for high school students learning English as a second language. Our strategy requires minimal effort in content preparation, one of the key challenges of alternatives such as home tutors or ITSs. We carried out a Randomized Controlled Trial (RCT) in four high-school classes, replacing traditional homework with GPT-4 homework sessions for the treatment group. We found that the treatment group had higher levels of satisfaction and desire to keep using the system among the students. This occurred without compromising learning outcomes, and one group even showed significantly better learning gains.

MAPLE: Enhancing Review Generation with Multi-Aspect Prompt LEarning in Explainable Recommendation

Ching-Wen Yang, Zhi-Quan Feng, Ying-Jia Lin, Che Wei Chen, Kun-da Wu, Hao Xu, Yao Jui-Feng, Hung-Yu Kao

Explainable Recommendation task is designed to receive a pair of user and item and output explanations to justify why an item is recommended to a user. Many models approach review generation as a proxy for explainable recommendations. While these models can produce fluent and grammatically correct sentences, they often lack preciseness and fail to provide personalized informative recommendations. To address this issue, we propose a personalized, aspect-controlled model called Multi-Aspect Prompt LEarner (MAPLE), which integrates aspect category as another input dimension to facilitate memorizing fine-grained aspect terms. Experiments conducted on two real-world review datasets in the restaurant domain demonstrate that MAPLE significantly outperforms baseline review-generation models. MAPLE excels in both text and feature diversity, ensuring that the generated content covers a wide range of aspects. Additionally, MAPLE delivers good generation quality while maintaining strong coherence and factual relevance. The code and dataset used in this paper can be found at https://github.com/Nana2929/MAPLE.

EducationQ: Evaluating LLMs' Teaching Capabilities Through Multi-Agent Dialogue Framework

Yao Shi, Rongkeng Liang, Yong Xu

Large Language Models (LLMs) increasingly serve as educational tools, yet evaluating their teaching capabilities remains challenging due to the resource-intensive, context-dependent, and methodologically complex nature of teacher-student interactions. We introduce EducationQ, a multi-agent dialogue framework that efficiently assesses teaching capabilities through simulated dynamic educational scenarios, featuring specialized agents for teaching, learning, and evaluation. Testing 14 LLMs across major AI Organizations (OpenAI, Meta, Google, Anthropic, and others) on 1,498 questions spanning 13 disciplines and 10 difficulty levels reveals that teaching effectiveness does not correlate linearly with model scale or general reasoning capabilities - with some smaller open-source models outperforming larger commercial counterparts in teaching contexts. This finding highlights a critical gap in current evaluations that prioritize knowledge recall over interactive pedagogy. Our mixed-methods evaluation, combining quantitative metrics with qualitative analysis and expert case studies, identifies distinct pedagogical strengths employed by top-performing models (e.g., sophisticated questioning strategies, adaptive feedback mechanisms). Human expert evaluations show 78% agreement with our automated qualitative analysis of effective teaching behaviors, validating our methodology. EducationQ demonstrates that LLMs-as-Teachers require specialized optimization beyond simple scaling, suggesting next-generation educational AI prioritize targeted enhancement of specific pedagogical effectiveness.

GRAM: Generative Recommendation via Semantic-aware Multi-granular Late Fusion

Sunkyung Lee, Minjin Choi, Eunseong Choi, Hye-young Kim, Jongwuk Lee

Generative recommendation is an emerging paradigm that leverages the extensive knowledge of large language models by formulating recommendations into a text-to-text generation task. However, existing studies face two key limitations in (i) incorporating implicit item relationships and (ii) utilizing rich yet lengthy item information. To address these challenges, we propose a Generative Recommender via semantic-Aware Multi-granular late fusion (GRAM), introducing two synergistic innovations. First, we design semantic-to-lexical translation to encode implicit hierarchical and collaborative item relationships into the vocabulary space of LLMs. Second, we present multi-granular late fusion to integrate rich semantics efficiently with minimal information loss. It employs separate encoders for multi-granular prompts, delaying the fusion until the decoding stage. Experiments on four benchmark datasets show that GRAM outperforms eight state-of-the-art generative recommendation models, achieving significant improvements of 11.5-16.0% in Recall@5 and 5.3-13.6% in NDCG@5. The source code is available at https://github.com/skleee/GRAM.

Session 2: Resources and Evaluation (Monday 28 July, 11:00-12:30, Hall 4/5)

EcomScriptBench: A Multi-task Benchmark for E-commerce Script Planning via Step-wise Intention-Driven Product Association

Weiqi Wang, Limeng Cui, Xin Liu, Sreyashi Nag, Wenju Xu, Chen Luo, Sheikh Muhammad Sarwar, Yang Li, Hansu Gu, Hui Liu, Changlong Yu, Jiaxin Bai, Yifan Gao, Haiyang Zhang, Qi He, Shuiwang Ji, Yangqiu Song

Goal-oriented script planning, or the ability to devise coherent sequences of actions toward specific goals, is commonly employed by humans to plan for typical activities. In e-commerce, customers increasingly seek LLM-based assistants to generate scripts and recommend products at each step, thereby facilitating convenient and efficient shopping experiences. However, this capability remains underexplored due to several challenges, including the inability of LLMs to simultaneously conduct script planning and product retrieval, difficulties in matching products caused by semantic discrepancies between planned actions and search queries, and a lack of methods and benchmark data for evaluation. In this paper, we step forward by formally defining the task of E-commerce Script Planning (EcomScript) as three sequential subtasks. We propose a novel framework that enables the scalable generation of product-enriched scripts by associating products with each step based on the semantic similarity between the actions and their purchase intentions. By applying our framework to real-world e-commerce data, we construct the very first large-scale EcomScript dataset, EcomScriptBench, which includes 605,229 scripts sourced from 2.4 million products. Human annotations are then conducted to provide gold labels for a sampled subset, forming an evaluation benchmark. Extensive experiments reveal that current (L)LMs face significant challenges with EcomScript tasks, even after fine-tuning, while injecting product purchase intentions improves their performance.

M-RewardBench: Evaluating Reward Models in Multilingual Settings

Srishti Gureja, Lester James Validad Miranda, Shayekh Bin Islam, Rishabh Maheshwary, Drishti Sharma, Gusti Triandi Winata, Nathan Lambert, Sebastian Ruder, Sara Hooker, Marzieh Fadaee

Reward models (RMs) have driven the state-of-the-art performance of LLMs today by enabling the integration of human feedback into the language modeling process. However, RMs are primarily trained and evaluated in English, and their capabilities in multilingual settings remain largely understudied. In this work, we conduct a systematic evaluation of several reward models in multilingual settings. We first construct the first-of-its-kind multilingual RM evaluation benchmark, M-RewardBench, consisting of 2.87k preference instances for 23 typologically diverse languages, that tests the chat, safety, reasoning, and translation capabilities of RMs. We then rigorously evaluate a wide range of reward models on M-RewardBench, offering fresh insights into their performance across diverse languages. We identify a significant gap in RMs' performances between English and non-English languages and show that RM preferences can change substantially from one language to another. We also present several findings on how different multilingual aspects impact RM performance. Specifically, we show that the performance of RMs is improved with improved translation quality. Similarly, we demonstrate that the models exhibit better performance for high-resource languages. We release M-RewardBench dataset and the codebase in this study to facilitate a better understanding of RM evaluation in multilingual settings.

CLEME2.0: Towards Interpretable Evaluation by Disentangling Edits for Grammatical Error Correction

Jingheng Ye, Zishan Xu, Yinghui Li, Linlin Song, Qingyu Zhou, Hai-Tao Zheng, Ying Shen, Wenhao Jiang, Hong-Gee Kim, Ruitong Liu, Xin Su, Zifei Shan

The paper focuses on the interpretability of Grammatical Error Correction (GEC) evaluation metrics, which received little attention in previous studies. To bridge the gap, we introduce **CLEME2.0**, a reference-based metric describing four fundamental aspects of GEC systems: hit-correction, wrong-correction, under-correction, and over-correction. They collectively contribute to exposing critical qualities and locating drawbacks of GEC systems. Evaluating systems by combining these aspects also leads to superior human consistency over other reference-based and reference-less metrics. Extensive experiments on two human judgment datasets and six reference datasets demonstrate the effectiveness and robustness of our method, achieving a new state-of-the-art result. Our codes are released at https://github.com/THUKElab/CLEME.

BelarusianGLUE: Towards a Natural Language Understanding Benchmark for Belarusian

Maksim Aparovich, Volha Harytskaya, Vladislav Poritski, Oksana Volchek, Pavel Smrz

In the epoch of multilingual large language models (LLMs), it is still challenging to evaluate the models' understanding of lower-resourced languages, which motivates further development of expert-crafted natural language understanding benchmarks. We introduce BelarusianGLUE — a natural language understanding benchmark for Belarusian, an East Slavic language, with 15K instances in five tasks: sentiment analysis, linguistic acceptability, word in context, Winograd schema challenge, textual entailment. A systematic evaluation of BERT models and LLMs against this novel benchmark reveals that both types of models approach human-level performance on easier tasks, such as sentiment analysis, but there is a significant gap in performance between machine and human on a harder task — Winograd schema challenge. We find the optimal choice of model type to be task-specific: e.g. BERT models underperform on textual entailment task but are competitive for linguistic acceptability. We release the datasets (https://hf.co/datasets/maaxap/BelarusianGLUE) and evaluation code (https://github.com/maaxap/BelarusianGLUE).

RuleArena: A Benchmark for Rule-Guided Reasoning with LLMs in Real-World Scenarios

Ruiwen Zhou, Wenyue Hua, Liangming Pan, Sitao Cheng, Xiaobao Wu, En Yu, William Yang Wang

This paper introduces RuleArena, a novel and challenging benchmark designed to evaluate the ability of large language models (LLMs) to follow complex, real-world rules in reasoning. Covering three practical domains -- airline baggage fees, NBA transactions, and tax regulations -- RuleArena assesses LLMs' proficiency in handling intricate natural language instructions that demand long-context understanding, logical reasoning, and accurate mathematical computation. Two key attributes distinguish RuleArena from traditional rule-based reasoning benchmarks: (1) it extends beyond standard first-order logic representations, and (2) it is grounded in authentic, practical scenarios, providing insights into the suitability and reliability of LLMs for real-world applications. Our findings reveal several notable limitations in LLMs: (1) they struggle to identify and apply the appropriate rules, frequently becoming confused by similar but distinct regulations, (2) they cannot consistently perform accurate mathematical computations, even when they correctly identify the relevant rules, and (3) in general, they perform poorly in the benchmark. We also observe a significant performance boost when LLMs are provided with external tools for oracle math and logic operations. These results highlight significant challenges and promising research directions in advancing LLMs' rule-guided reasoning capabilities in real-life applications. Our codes and data are publicly available on https://github.com/skyriver-2000/rulearena.

JuStRank: Benchmarking LLM Judges for System Ranking

Ariel Gera, Odellia Boni, Yotam Perlitz, Roy Bar-Haim, Lilach Eden, Asaf Yehudai

Given the rapid progress of generative AI, there is a pressing need to systematically compare and choose between the numerous models and configurations available. The scale and versatility of such evaluations make the use of LLM-based judges a compelling solution for this challenge. Crucially, this approach requires first to validate the quality of the LLM judge itself. Previous work has focused on instance-based assessment of LLM judges, where a judge is evaluated over a set of responses, or response pairs, while being agnostic to their source systems. We argue that this setting overlooks critical factors affecting system-level ranking, such as a judge's positive or negative bias towards certain systems. To address this gap, we conduct the first large-scale study of LLM judges as system rankers. System scores are generated by aggregating judgment scores over multiple system outputs, and the judge's quality is assessed by comparing the resulting system ranking to a human-based ranking. Beyond overall judge assessment, our analysis provides a fine-grained characterization of judge behavior, including their decisiveness and bias.

EscapeBench: Towards Advancing Creative Intelligence of Language Model Agents

Cheng Qian, Peixuan Han, Qinyu Luo, Bingxiang He, Xiusi Chen, Yuji Zhang, Hongyi Du, Jiarui Yao, Xiaocheng Yang, Denghui Zhang, Yunzhu Li, Heng Ji

Language model agents excel in long-session planning and reasoning, but existing benchmarks primarily focus on goal-oriented tasks with explicit objectives, neglecting creative adaptation in unfamiliar environments. To address this, we introduce EscapeBench—a benchmark suite of room escape game environments designed to challenge agents with creative reasoning, unconventional tool use, and iterative problem-solving to uncover implicit goals. Our results show that current LM models, despite employing working memory and Chain-of-Thought reasoning, achieve only 15% average progress without hints, highlighting their limitations in creativity. To bridge this gap, we propose EscapeAgent, a framework designed to enhance creative reasoning through Foresight (innovative tool use) and Reflection (identifying unsolved tasks). Experiments show that EscapeAgent can execute action chains over 1,000 steps while maintaining logical coherence. It navigates and completes games with up to 40% fewer steps and hints, performs robustly across difficulty levels, and achieves higher action success rates with more efficient and innovative puzzle-solving strategies.

Aligning AI Research with the Needs of Clinical Coding Workflows: Eight Recommendations Based on US Data Analysis and Critical Review

Yidong Gan, Maciej Rybinski, Ben Hachey, Jonathan K. Kummerfeld

Clinical coding is crucial for healthcare billing and data analysis. Manual clinical coding is labour-intensive and error-prone, which has motivated research towards full automation of the process. However, our analysis, based on US English electronic health records and automated coding research using these records, shows that widely used evaluation methods are not aligned with real clinical contexts. For example, evaluations that focus on the top 50 most common codes are an oversimplification, as there are thousands of codes used in practice. This position paper aims to align AI coding research more closely with practical challenges of clinical coding. Based on our analysis, we offer eight specific recommendations, suggesting ways to improve current evaluation methods. Additionally, we propose new AI-based methods beyond automated coding, suggesting alternative approaches to assist clinical coders in their workflows.

MARS: Benchmarking the Metaphysical Reasoning Abilities of Language Models with a Multi-task Evaluation Dataset

Weiqi Wang, Yangqiu Song

To enable Large Language Models (LLMs) to function as conscious agents with generalizable reasoning capabilities, it is crucial that they possess the ability to ***comprehend situational changes (transitions) in distribution*** triggered by environmental factors or actions from other agents. Despite its fundamental significance, this ability remains underexplored due to the complexity of modeling infinite possible changes in an event and their associated distributions, coupled with the lack of benchmark data with situational transitions. Addressing these gaps, we propose a novel formulation of ***reasoning with distributional changes as a three-step discriminative process***, termed as ***MetAphysical ReaSoning***. We then introduce the first-ever benchmark, **MARS**, comprising three tasks corresponding to each step. These tasks systematically assess LLMs' capabilities in reasoning the plausibility of (i) changes in actions, (ii) states caused by changed actions, and (iii) situational transitions driven by changes in action. Extensive evaluations with 20 (L)LMs of varying sizes and methods indicate that all three tasks in this process pose significant challenges, even after fine-tuning. Further analyses reveal potential causes for the underperformance of LLMs and demonstrate that pre-training on large-scale conceptualization taxonomies can potentially enhance LMs' metaphysical reasoning capabilities. Our data and models are publicly accessible at https://github.com/HKUST-KnowComp/MARS.

TUNA: Comprehensive Fine-grained Temporal Understanding Evaluation on Dense Dynamic Videos

Fanheng Kong, Jingyuan Zhang, Hongzhi Zhang, Shi Feng, Daling Wang, Linhao Yu, Xingguang Ji, Yu Tian, V. W., Fuzheng Zhang

Videos are unique in their integration of temporal elements, including camera, scene, action, and attribute, along with their dynamic relationships over time. However, existing benchmarks for video understanding often treat these properties separately or narrowly focus on specific aspects, overlooking the holistic nature of video content. To address this, we introduce TUNA, a temporal-oriented benchmark for fine-grained understanding on dense dynamic videos, with two complementary tasks: captioning and QA. Our TUNA features diverse video scenarios and dynamics, assisted by interpretable and robust evaluation criteria. We evaluate several leading models on our benchmark, providing fine-grained performance assessments across various dimensions. This evaluation reveals key challenges in video temporal understanding, such as limited action description, inadequate multi-subject understanding, and insensitivity to camera motion, offering valuable insights for improving video understanding models.

UnSeenTimeQA: Time-Sensitive Question-Answering Beyond LLMs' Memorization

Md Nayem Uddin, Amir Saeidi, Divij Handa, Agastya Seth, Tran Cao Son, Eduardo Blanco, Steven Corman, Chitta Baral

This paper introduces UnSeenTimeQA, a novel data contamination-free time-sensitive question-answering (TSQA) benchmark. It differs from existing TSQA benchmarks by avoiding web-searchable queries grounded in the real world. We present a series of time-sensitive event scenarios based on synthetically generated facts. It requires large language models (LLMs) to engage in genuine temporal reasoning without depending on the factual knowledge acquired during the pre-training phase. Our data generation framework enables on-demand generation of new samples, mitigating the risk of data leakage. We designed three types of time-sensitive questions to test LLMs' temporal reasoning abilities over sequential and parallel event occurrences. Our evaluation of five LLMs on synthetic fact-based TSQA reveals mixed results: while they perform well on simpler subsets, their overall performance remains inferior as compared to real world fact-based TSQA. Error analysis indicates that LLMs face difficulties in reasoning over long-range event dependencies and parallel events.

LLäMmlein : Transparent, Compact and Competitive German-Only Language Models from Scratch

Jan Pfister, Julia Wunderle, Andreas Hotho

We transparently create two German-only decoder models, LLäMmlein 120M and 1B, from scratch and publish them, along with the training data, for the (German) NLP research community to use. The model training involved several key steps, including data preprocessing/filtering, the creation of a German tokenizer, the training itself, as well as the evaluation of the final models on various benchmarks, also against existing models. Throughout the training process, multiple checkpoints were saved in equal intervals and analyzed using the German SuperGLEBer benchmark to gain insights into the models' learning process. Compared to state-of-the-art models on the SuperGLEBer benchmark, both LLäMmlein models performed competitively, consistently matching or surpassing models with similar parameter sizes. The results show that the models' quality scales with size as expected, but performance improvements on some tasks plateaued early during training, offering valuable insights into resource allocation for future models.

LegalAgentBench: Evaluating LLM Agents in Legal Domain

Haitao Li, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin, Yueyue WU, Guozhi Yuan, Yiran HU, Wuyue Wang, Yiqun LIU, Minlie Huang

With the increasing intelligence and autonomy of LLM Agents, their potential applications in the legal domain are becoming increasingly apparent. However, existing general-domain benchmarks are unable to fully capture the complexity and subtle nuances inherent in real-world judicial cognition and decision-making. Therefore, we propose LegalAgentBench, a comprehensive benchmark specifically designed to evaluate LLM Agents in the Chinese legal domain. LegalAgentBench includes 17 corpora from real-world legal scenarios and provides 37 tools for interacting with external knowledge. To cover tasks of varying difficulty and types, we designed a scalable task construction process that enables a more precise evaluation of performance in both tool utilization and reasoning. Moreover, Beyond assessing performance through the success rate of final outcomes, LegalAgentBench incorporates keyword analysis during intermediate processes to calculate progress rates, facilitating a more fine-grained evaluation. We evaluated eight popular LLMs, highlighting the strengths, limitations, and potential areas for improvement of existing models and methods. LegalAgentBench sets a new benchmark for the practical application of LLMs in the legal domain, with its code and data available at https://github.com/CSHaitao/LegalAgentBench.

Behind Closed Words: Creating and Investigating the forePLay Annotated Dataset for Polish Erotic Discourse

Anna Kołos, Katarzyna Lorenc, Emilia Wiśnios, Agnieszka Karlińska

The surge in online content has created an urgent demand for robust detection systems, especially in non-English contexts where current tools demonstrate significant limitations. We introduce forePLay, a novel Polish-language dataset for erotic content detection, comprising over 24,000 annotated sentences. The dataset features a multidimensional taxonomy that captures ambiguity, violence, and socially unacceptable behaviors. Our comprehensive evaluation demonstrates that specialized Polish language models achieve superior performance compared to multilingual alternatives, with transformer-based architectures showing particular strength in handling imbalanced categories. The dataset and accompanying analysis establish essential frameworks for developing linguistically-aware content moderation systems, while highlighting critical considerations for extending such capabilities to morphologically complex languages.

GigaSpeech 2: An Evolving, Large-Scale and Multi-domain ASR Corpus for Low-Resource Languages with Automated Crawling, Transcription and Refinement

Yifan Yang, Zheshu Song, Jianheng Zhuo, Mingyu Cui, Jinpeng Li, Bo Yang, Yexing Du, Ziyang Ma, Xunying Liu, Ziyuan Wang, Ke Li, Shuai Fan, Kai Yu, Wei-Qiang Zhang, Guoguo Chen, Xie Chen

The evolution of speech technology has been spurred by the rapid increase in dataset sizes. Traditional speech models generally depend on a large amount of labeled training data, which is scarce for low-resource languages. This paper presents GigaSpeech 2, a large-scale, multi-domain, multilingual speech recognition corpus. It is designed for low-resource languages and does not rely on paired speech and text data. GigaSpeech 2 comprises about 30,000 hours of automatically transcribed speech, including Thai, Indonesian, and Vietnamese, gathered from unlabeled YouTube videos. We also introduce an automated pipeline for data crawling, transcription, and label refinement. Specifically, this pipeline involves Whisper for initial transcription, MMS for forced alignment, and multi-dimensional filtering for data quality assurance. A modified Noisy Student Training is developed to further refine flawed pseudo labels iteratively, thereby enhancing model performance. Experimental results on our manually transcribed evaluation set and two public test sets from Common Voice and FLEURS confirm our corpus's high quality and broad applicability. Notably, ASR models trained on GigaSpeech 2 can reduce the word error rate for Thai, Indonesian, and Vietnamese on our challenging and realistic YouTube test set by 25% to 40% compared to Whisper large-v3, with merely 10% model parameters. Furthermore, our ASR models trained on GigaSpeech 2 yield superior performance compared to commercial services. We hope that our newly introduced corpus and pipeline will open a new avenue for low-resource speech recognition and significantly facilitate research in this area.

TRACT: Regression-Aware Fine-tuning Meets Chain-of-Thought Reasoning for LLM-as-a-Judge

Cheng-Han Chiang, Hung-yi Lee, Michal Lukasik

The LLM-as-a-judge paradigm uses large language models (LLMs) for automated text evaluation, assigning a score to the input based on scoring rubrics. Existing methods for fine-tuning LLM-as-a-judge use cross-entropy (CE) loss, which neglects the numeric nature of score prediction. Recent work addresses numerical prediction limitations of LLM fine-tuning through regression-aware fine-tuning but does not consider chain-of-thought (CoT) reasoning for score prediction. In this paper, we introduce TRACT (Two-stage Regression-Aware fine-tuning with CoT), which combines CoT reasoning with regression-aware training. TRACT uses a two-stage process: first, it fine-tunes the seed LLM to generate CoTs, which serve as the training data for the second stage; next, it uses these self-generated CoTs to retrain the seed LLM. The fine-tuning objective of TRACT applies CE loss for CoT reasoning and regression-aware loss for the score. Experiments across four LLM-as-a-judge datasets and two LLMs show that TRACT significantly outperforms existing methods. Extensive ablation studies validate the effectiveness of each component in TRACT.

ToolHop: A Query-Driven Benchmark for Evaluating Large Language Models in Multi-Hop Tool Use

Junjie Ye, Zhengyin Du, Xuesong Yao, Weijian Lin, Yufei Xu, Zehui Chen, Zaiyuan Wang, Sining Zhu, Zhiheng Xi, Siyu Yuan, Tao Gui, Qi Zhang, Xuanjing Huang, Jiecao Chen

Effective evaluation of multi-hop tool use is critical for analyzing the understanding, reasoning, and function-calling capabilities of large language models (LLMs). However, progress has been hindered by a lack of reliable evaluation datasets. To address this, we present ToolHop, a dataset comprising 995 user queries and 3,912 associated tools, specifically designed for rigorous evaluation of multi-hop tool use. ToolHop ensures diverse queries, meaningful interdependencies, locally executable tools, detailed feedback, and verifiable answers through a novel query-driven data construction approach that includes tool creation, document refinement, and code generation. We evaluate 14 LLMs across five model families (i.e., LLaMA3.1, Qwen2.5, Gemini1.5, Claude3.5, and GPT), uncovering significant challenges in handling multi-hop tool-use scenarios. The leading model, GPT-4o, achieves an accuracy of 49.04%, underscoring substantial room for improvement. Further analysis reveals variations in tool-use strategies for various families, offering actionable insights to guide the development of more effective approaches. Code and data can be found in https://huggingface.co/datasets/bytedance-research/ToolHop.

A-TASC: Asian TED-Based Automatic Subtitling Corpus

Yuhan Zhou, Naoki Yoshinaga

Subtitles play a crucial role in improving the accessibility of the vast amount of audiovisual content available on the Internet, allowing audiences worldwide to comprehend and engage with this content in various languages. Automatic subtitling (AS) systems are essential for alleviating the substantial workload of human transcribers and translators. However, existing AS corpora and the primary metric SubER focus on European languages. This paper introduces A-TASC, an Asian TED-based automatic subtitling corpus derived from English TED Talks, comprising nearly 800 hours of audio segments, aligned English transcripts, and subtitles in Chinese, Japanese, Korean, and Vietnamese. We then present SacreSubER, a modification of SubER, to enable the reliable evaluation of subtitle quality for languages without explicit word boundaries. Experimental results, using both end-to-end systems and pipeline approaches built on strong ASR and LLM components, validate the quality of the proposed corpus and reveal differences in AS performance between European and Asian languages. The code to build our corpus is released.

LazyReview A Dataset for Uncovering Lazy Thinking in NLP Peer Reviews

Sukannya Purkayastha, Zhuang Li, Anne Lauscher, Lizhen Qu, Iryna Gurevych

Peer review is a cornerstone of quality control in scientific publishing. With the increasing workload, the unintended use of `quick' heuristics, referred to as \emph{lazy thinking}, has emerged as a recurring issue compromising review quality. Automated methods to detect such heuristics can help improve the peer-reviewing process. However, there is limited NLP research on this issue, and no real-world dataset exists to support the development of detection tools. This work introduces \textsc{LazyReview}, a dataset of peer-review sentences annotated with fine-grained \emph{lazy thinking} categories. Our analysis reveals that Large Language Models (LLMs) struggle to detect these instances in a zero-shot setting. However, instruction-based fine-tuning on our dataset significantly boosts performance by 10-20 performance points, highlighting the importance of high-quality training data. Furthermore, a controlled experiment demonstrates that reviews revised with \emph{lazy thinking} feedback are more comprehensive and actionable than those written without such feedback. We will release our dataset and the enhanced guidelines that can be used to train junior reviewers in the community.

LongBench v2: Towards Deeper Understanding and Reasoning on Realistic Long-context Multitasks

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, Juanzi Li

This paper introduces LongBench v2, a benchmark designed to assess the ability of LLMs to handle long-context problems requiring deep understanding and reasoning across real-world multitasks. LongBench v2 consists of 503 challenging multiple-choice questions, with contexts ranging from 8k to 2M words, across six major task categories: single-document QA, multi-document QA, long in-context learning, long-dialogue history understanding, code repository understanding, and long structured data understanding. To ensure the breadth and the practicality, we collect data from nearly 100 highly educated individuals with diverse professional backgrounds. We employ both automated and manual review processes to maintain high quality and difficulty, resulting in human experts achieving only 53.7% accuracy under a 15-minute time constraint. Our evaluation reveals that the best-performing model, when directly answers the questions, achieves only 50.1% accuracy. In contrast, the o1-preview model, which includes longer reasoning, achieves 57.7%, surpassing the human baseline by 4%. These results highlight the importance of enhanced reasoning ability and scaling inference-time compute to tackle the long-context challenges in LongBench v2.

Establishing Trustworthy LLM Evaluation via Shortcut Neuron Analysis

Kejian Zhu, Shangqing Tu, Zhuoran Jin, Lei Hou, Juanzi Li, Jun Zhao

The development of large language models (LLMs) depends on **trustworthy evaluation**. However, most current evaluations rely on public benchmarks, which are prone to data contamination issues that significantly compromise fairness. Previous researches have focused on constructing dynamic benchmarks to address contamination. However, continuously building new benchmarks is costly and cyclical. In this work, we aim to tackle contamination by analyzing the mechanisms of contaminated models themselves. Through our experiments, we discover that the overestimation of contaminated models is likely due to parameters acquiring shortcut solutions in training. We further propose a novel method for identifying shortcut neurons through **comparative and causal analysis**. Building on this, we introduce an evaluation method called **shortcut neuron patching** to suppress shortcut neurons. Experiments validate the effectiveness of our approach in mitigating contamination. Additionally, our evaluation results exhibit a strong linear correlation with MixEval, a recently released trustworthy benchmark, achieving a Spearman coefficient (\rho) exceeding 0.95. This high correlation indicates that our method closely reveals true capabilities of the models and is trustworthy. We conduct further experiments to demonstrate the generalizability of our method across various benchmarks and hyperparameter settings. **Code**: https://github.com/GaryStack/Trustworthy-Evaluation.

Can LLMs Simulate L2-English Dialogue? An Information-Theoretic Analysis of L1-Dependent Biases

Rena Wei Gao, Xuetong Wu, Tatsuki Kuribayashi, Mingrui Ye, Siya Qi, Carsten Roever, Yuanxing Liu, Zheng Yuan, Jey Han Lau

This study evaluates Large Language Models' (LLMs) ability to simulate non-native-like English use observed in human second language (L2) learners interfered with by their native first language (L1). In dialogue-based interviews, we prompt LLMs to mimic L2 English learners with specific L1s (e.g., Japanese, Thai, Urdu) across seven languages, comparing their outputs to real L2 learner data. Our analysis examines L1-driven linguistic biases, such as reference word usage and avoidance behaviors, using information-theoretic and distributional density measures. Results show that modern LLMs (e.g., Qwen2.5, LLAMA3, DeepseekV3, GPT 4o) replicate L1-dependent patterns observed in human L2 data, with distinct influences from various languages (e.g., Japanese, Korean, and Mandarin significantly affect tense agreement, and Urdu influences noun-verb collocations). Our results reveal LLMs' potential for L2 dialogue generation and evaluation for future educational applications.

Auto-Arena: Automating LLM Evaluations with Agent Peer Battles and Committee Discussions

Ruochen Zhao, Wenxuan Zhang, Yew Ken Chia, Weiwen Xu, Deli Zhao, Lidong Bing

As LLMs continuously evolve, there is an urgent need for a reliable evaluation method that delivers trustworthy results promptly. Currently, static benchmarks suffer from inflexibility and unreliability, leading users to prefer human voting platforms like Chatbot Arena. However, human evaluations require significant manual effort. Therefore, we propose Auto-Arena, an innovative framework that automates the entire evaluation process using LLM-powered agents. Firstly, an LLM examiner generates questions. Then, two LLM candidates engage in a multi-round peer battle based on the questions, aiming at revealing their true performance differences. Finally, a committee of LLM judges collaboratively discusses and decides the winner, reducing bias and enhancing fairness. During the peer battles, we observe intriguing scenarios where the LLM candidates display competitive behaviors and learn from the opponents. In our extensive experiments involving 15 recent LLMs, Auto-Arena shows a 92.14% correlation with human preferences, surpassing all previous expert-annotated benchmarks without any manual efforts. Auto-Arena offers a promising alternative to current human evaluation platforms for evaluating LLMs automatically.

How Humans and LLMs Organize Conceptual Knowledge: Exploring Subordinate Categories in Italian

Andrea Pedrotti, Giulia Rambelli, Caterina Villani, Marianna Bolognesi

People can categorize the same entity at multiple taxonomic levels, such as basic (bear), superordinate (animal), and subordinate (grizzly bear). While prior research has focused on basic-level categories, this study is the first attempt to examine the organization of categories by analyzing exemplars produced at the subordinate level. We present a new Italian psycholinguistic dataset of human-generated exemplars for 187 concrete words. We then leverage these data to evaluate whether textual and vision LLMs produce meaningful exemplars that align with human category organization across three key tasks: exemplar generation, category induction, and typicality judgment. Our findings show a low alignment between humans and LLMs, consistent with previous studies. However, their performance varies notably across different semantic domains. Ultimately, this study highlights both the promises and the constraints of using AI-generated exemplars to support psychological and linguistic research.

On the Limit of Language Models as Planning Formalizers

Cassie Huang, Li Zhang

Large Language Models have been found to create plans that are neither executable nor verifiable in grounded environments. An emerging line of work demonstrates success in using the LLM as a formalizer to generate a formal representation of the planning domain in some language, such as Planning Domain Definition Language (PDDL). This formal representation can be deterministically solved to find a plan. We systematically evaluate this methodology while bridging some major gaps. While previous work only generates a partial PDDL representation, given templated, and therefore unrealistic environment descriptions, we generate the complete representation given descriptions of various naturalness levels. Among an array of observations critical to improve LLMs' formal planning abilities, we note that most large enough models can effectively formalize descriptions as PDDL, outperforming those directly generating plans, while being robust to lexical perturbation. As the descriptions become more natural-sounding, we observe a decrease in performance and provide detailed error analysis.

On the Robustness of RAG Systems in Educational Question Answering under Knowledge Discrepancies

Tianshi Zheng, Weihan Li, Jiaxin Bai, Weiqi Wang, Yangqiu Song

Retrieval-Augmented Generation (RAG) systems show remarkable potential as question answering tools in the K-12 Education domain, where knowledge is typically queried within the restricted scope of authoritative textbooks. However, discrepancies between these textbooks and the parametric knowledge inherent in Large Language Models (LLMs) can undermine the effectiveness of RAG systems. To systematically investigate RAG system robustness against such knowledge discrepancies, we introduce KnowShiftQA. This novel question answering dataset simulates these discrepancies by applying deliberate hypothetical knowledge updates to both answers and source documents, reflecting how textbook knowledge can shift. KnowShiftQA comprises 3,005 questions across five subjects, designed with a comprehensive question typology focusing on context utilization and knowledge integration. Our extensive experiments on retrieval and question answering performance reveal that most RAG systems suffer a substantial performance drop when faced with these knowledge discrepancies. Furthermore, questions requiring the integration of contextual (textbook) knowledge with parametric (LLM) knowledge pose a significant challenge to current LLMs.

People who frequently use ChatGPT for writing tasks are accurate and robust detectors of AI-generated text

Jenna Russell, Marzena Karpinska, Mohit Iyyer

In this paper, we study how well humans can detect text generated by commercial LLMs (GPT-4o, Claude, o1). We hire annotators to read 300 non-fiction English articles, label them as either human-written or AI-generated, and provide paragraph-length explanations for their decisions. Our experiments show that annotators who frequently use LLMs for writing tasks excel at detecting AI-generated text, even without any specialized training or feedback. In fact, the majority vote among five such "expert" annotators misclassifies only 1 of 300 articles, significantly outperforming most commercial and open-source detectors we evaluated even in the presence of evasion tactics like paraphrasing and humanization. Qualitative analysis of the experts' free-form explanations shows that while they rely heavily on specific lexical clues ('AI vocabulary'), they also pick up on more complex phenomena within the text (e.g., formality, originality, clarity) that are challenging to assess for automatic detectors. We release our annotated dataset and code to spur future research into both human and automated detection of AI-generated text.

What Makes a Good Natural Language Prompt?

Do Xuan Long, Duy Dinh, Ngoc-Hai Nguyen, Kenji Kawaguchi, Nancy F. Chen, Shafiq Joty, Min-Yen Kan

As large language models (LLMs) have progressed towards more human-like and human--AI communications prevalent, prompting has emerged as a decisive component. However, there is limited conceptual consensus on what exactly quantifies natural language prompts. We attempt to address this question by conducting a meta-analysis surveying 150+ prompting-related papers from leading NLP and AI conferences (2022–2024), and blogs. We propose a property- and human-centric framework for evaluating prompt quality, encompassing 21 properties categorized into six dimensions. We then examine how existing studies assess their impact on LLMs, revealing their imbalanced support across models and tasks, and substantial research gaps. Further, we analyze correlations among properties in high-quality natural language prompts, deriving prompting recommendations. Finally, we explore multi-property prompt enhancements in reasoning tasks, observing that single-property enhancements often have the greatest impact. Our findings establish a foundation for property-centric prompt evaluation and optimization, bridging the gaps between human--AI communication and opening new prompting research directions.

LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, Alberto Testoni

There is an increasing trend towards evaluating NLP models with LLMs instead of human judgments, raising questions about the validity of these evaluations, as well as their reproducibility in the case of proprietary models. We provide JUDGE-BENCH, an extensible collection of 20 NLP datasets with human annotations covering a broad range of evaluated properties and types of data, and comprehensively evaluate 11 current LLMs, covering both open-weight and proprietary models, for their ability to replicate the annotations. Our evaluations show substantial variance across models and datasets. Models are reliable evaluators on some tasks, but overall display substantial variability depending on the property being evaluated, the expertise level of the human judges, and whether the language is human or model-generated. We conclude that LLMs should be carefully validated against human judgments before being used as evaluators.

LexGen: Domain-aware Multilingual Lexicon Generation

Ayush Maheshwari, Atul Kumar Singh, N J Karthika, Krishnakant Bhatt, Preethi Jyothi, Ganesh Ramakrishnan

Lexicon or dictionary generation across domains has the potential for societal impact, as it can potentially enhance information accessibility for a diverse user base while preserving language identity. Prior work in the field primarily focuses on bilingual lexical induction, which deals with word alignments using mapping-based or corpora-based approaches. However, these approaches do not cater to domain-specific lexicon generation that consists of domain-specific terminology. This task becomes particularly important in specialized medical, engineering, and other technical domains, owing to the highly infrequent usage of the terms and scarcity of data involving domain-specific terms especially for low-resource languages. We propose a new model to generate dictionary words for 6 Indian languages in the multi-domain setting. Our model consists of domain-specific and domain-generic layers that encode information, and these layers are invoked via a learnable routing technique. We also release a new benchmark dataset consisting of >75K translation pairs across 6 Indian languages spanning 8 diverse domains. We conduct both zero-shot and few-shot experiments across multiple domains to show the efficacy of our proposed model in generalizing to unseen domains and unseen languages. Additionally, we also perform a human post-hoc evaluation on unseen languages. The source code and dataset is present at https://github.com/Atulkmrsingh/lexgen.

CKnowEdit: A New Chinese Knowledge Editing Dataset for Linguistics, Facts, and Logic Error Correction in LLMs

Jizhan Fang, Tianhe Lu, Yunzhi Yao, Ziyan Jiang, Xin Xu, Huajun Chen, Ningyu Zhang

Chinese, as a linguistic system rich in depth and complexity, is characterized by distinctive elements such as ancient poetry, proverbs, idioms, and other cultural constructs. However, current Large Language Models (LLMs) face limitations in these specialized domains, highlighting the need for the development of comprehensive datasets that can assess, continuously update, and progressively improve these culturally-grounded linguistic competencies through targeted training optimizations. To address this gap, we introduce CKnowEdit, the first-ever Chinese knowledge editing dataset designed to correct linguistic, factual, and logical errors in LLMs. We collect seven types of knowledge from a wide range of sources, including classical texts, idioms, and content from Baidu Tieba Ruozhiba, taking into account the unique polyphony, antithesis, and logical structures inherent in the Chinese language. By analyzing this dataset, we highlight the challenges current LLMs face in mastering Chinese. Furthermore, our evaluation of state-of-the-art knowledge editing techniques reveals opportunities to advance the correction of Chinese knowledge.

Sample-Efficient Human Evaluation of Large Language Models via Maximum Discrepancy Competition

Kehua Feng, Keyan Ding, tan hongzhi, Kede Ma, Zhihua Wang, Shuangquan Guo, Cheng yuzhou, Ge Sun, Guozhou Zheng, Qiang Zhang, Huajun Chen

The past years have witnessed a proliferation of large language models (LLMs). Yet, reliable evaluation of LLMs is challenging due to the inaccuracy of standard metrics in human perception of text quality and the inefficiency in sampling informative test examples for human evaluation. This paper presents a sample-efficient human evaluation method for LLMs based on the principle of MAximum Discrepancy (MAD) competition. MAD automatically selects a small set of informative input instructions, each of which maximizes the discrepancy of two LLMs' reponses, which are subsequently subject to three-alternative forced choice by human subjects. The pairwise comparison results of multiple LLMs are then aggregated into a global ranking using the Elo rating system. We compare eight representative LLMs in terms of four skills: knowledge understanding, mathematical reasoning, writing, and coding. Experimental results show that the proposed method reliably achieves the "golden" ranking of LLMs with a minimum set of input instructions, which in turn reveal their relative strengths and weaknesses, and offers valuable insights for further LLM advancement.

CaLMQA: Exploring culturally specific long-form question answering across 23 languages

Shane Arora, Marzena Karpinska, Hung-Ting Chen, Ipsita Bhattacharjee, Mohit Iyyer, Eunsol Choi

Despite rising global usage of large language models (LLMs), their ability to generate *long-form* answers to *culturally specific* questions remains unexplored in many languages. To fill this gap, we perform the first study of textual multilingual long-form QA by creating CaLMQA, a dataset of **51.7K** culturally specific questions across **23** different languages. We define culturally specific questions as those that refer to concepts unique to one or a few cultures, or have different answers depending on the cultural or regional context. We obtain these questions by crawling naturally-occurring questions from community web forums in high-resource languages, and by hiring native speakers to write questions in under-resourced, rarely-studied languages such as Fijian and Kirundi. Our data collection methodologies are translation-free, enabling the collection of culturally unique questions like "Kuber iki umwami wa mbere w'uburundi yitwa Ntare?" (Kirundi; English translation: "Why was the first king of Burundi called Ntare (Lion)?"). We evaluate factuality, relevance and surface-level quality of LLM-generated long-form answers, finding that (1) for many languages, even the best models make critical surface-level errors (e.g., answering in the wrong language, repetition), especially for low-resource languages; and (2) answers to culturally specific questions contain more factual errors than answers to culturally agnostic questions – questions that have consistent meaning and answer across many cultures. We release CaLMQA to facilitate future research in cultural and multilingual long-form QA.

GODBench: A Benchmark for Multimodal Large Language Models in Video Comment Art

Chenkai Zhang, Yiming Lei, Zeming Liu, Haitao Leng, ShaoGuo Liu, Tingting Gao, Qingjie Liu, Yunhong Wang

Video Comment Art enhances user engagement by providing creative content that conveys humor, satire, or emotional resonance, requiring a nuanced and comprehensive grasp of cultural and contextual subtleties. Although Multimodal Large Language Models (MLLMs) and Chain-of-Thought (CoT) have demonstrated strong reasoning abilities in STEM tasks (e.g. mathematics and coding), they still struggle to generate creative expressions such as resonant jokes and insightful satire. Moreover, existing benchmarks are constrained by their limited modalities and insufficient categories, hindering the exploration of comprehensive creativity in video-based Comment Art creation. To address these limitations, we introduce **GODBench**, a novel benchmark that integrates video and text modalities to systematically evaluate MLLMs' abilities to compose Comment Art. Furthermore, inspired by the propagation patterns of waves in physics, we propose **Ripple of Thought (RoT)**, a multi-step reasoning framework designed to enhance the creativity of MLLMs. Extensive experiments on GODBench reveal that existing MLLMs and CoT methods still face significant challenges in understanding and generating creative video comments. In contrast, RoT provides an effective approach to improving creative composing, highlighting its potential to drive meaningful advancements in MLLM-based creativity.

AbGen: Evaluating Large Language Models in Ablation Study Design and Evaluation for Scientific Research

Yilun Zhao, Weiyuan Chen, Zhijian Xu, Yixin Liu, Chengye Wang, Manasi Patwardhan, Lovekesh Vig, Arman Cohan

We introduce AbGen, the first benchmark designed to evaluate the capabilities of LLMs in designing ablation studies for scientific research. AbGen consists of 2,000 expert-annotated examples derived from 677 NLP papers. In this benchmark, LLMs are tasked with generating detailed ablation study designs for a specified module or process based on the given research context. Our evaluation of leading LLMs, such as GPT-4o and Llama-3.1, highlights a significant performance gap between these models and human experts in terms of the importance, faithfulness, and soundness of the ablation study designs. Moreover, we demonstrate that current automated evaluation methods are not reliable for our task, as they show a significant discrepancy when compared to human assessment. To better investigate this, we develop AbGen-Eval, a meta-evaluation benchmark designed to assess the reliability of commonly used automated evaluation systems in measuring LLM performance on our task. We investigate various LLM-based evaluation methods on AbGen-Eval, providing insights for future research on developing more effective and reliable LLM-based evaluation systems for complex scientific tasks.

MMLU-CF: A Contamination-free Multi-task Language Understanding Benchmark

Qihao Zhao, Yangyu Huang, Tengchao Lv, Lei Cui, Qinzheng Sun, Shaoguang Mao, Xin Zhang, Ying Xin, Qiufeng Yin, Scarlett Li, Furu Wei

Multiple-choice question (MCQ) datasets like Massive Multitask Language Understanding (MMLU) are widely used to evaluate the commonsense, understanding, and problem-solving abilities of large language models (LLMs). However, the open-source nature of these benchmarks and the broad sources of training data for LLMs have inevitably led to benchmark contamination, resulting in unreliable evaluation. To alleviate this issue, we propose the contamination-free MCQ benchmark called MMLU-CF, which reassesses LLMs' understanding of world knowledge by averting both unintentional and malicious data contamination. To mitigate unintentional data contamination, we source questions from a broader domain of over 200 billion webpages and apply three specifically designed decontamination rules. To prevent malicious data contamination, we divide the benchmark into validation and test sets with similar difficulty and subject distributions. The test set remains closed-source to ensure reliable results, while the validation set is publicly available to promote transparency and facilitate independent evaluation. The performance gap between these two sets of LLMs will indicate the contamination degree on the validation set in the future. We evaluated over 40 mainstream LLMs on the MMLU-CF. Compared to the original MMLU, not only LLMs' performances significantly dropped but also the performance rankings of them changed considerably. This indicates the effectiveness of our approach in establishing a contamination-free and fairer evaluation standard.

Enabling LLM Knowledge Analysis via Extensive Materialization

Yujia Hu, Tuan-Phong Nguyen, Shrestha Ghosh, Simon Razniewski

Large language models (LLMs) have majorly advanced NLP and AI, and next to their ability to perform a wide range of procedural tasks, a major success factor is their internalized factual knowledge. Since (Petroni et al., 2019), analyzing this knowledge has gained attention. However, most approaches investigate one question at a time via modest-sized pre-defined samples, introducing an ``availability bias'' (Tverski and Kahnemann, 1973) that prevents the analysis of knowledge (or beliefs) of LLMs beyond the experimenter's predisposition. To address this challenge, we propose a novel methodology to comprehensively materialize an LLM's factual knowledge through recursive querying and result consolidation. Our approach is a milestone for LLM research, for the first time providing constructive insights into the scope and structure of LLM knowledge (or beliefs). As a prototype, we extract a knowledge base (KB) comprising 101 million relational triples for over 2.9 million entities from GPT-4o-mini. We use GPTKB to exemplarily analyze GPT-4o-mini's factual knowledge in terms of scale, accuracy, bias, cutoff and consistency, at the same time. Our resource is accessible at https://gptkb.org.

Value Portrait: Understanding Values of LLMs with Human-aligned Benchmark

Jongwook Han, Dongmin Choi, Woojung Song, Eun-Ju Lee, Yohan Jo

The importance of benchmarks for assessing the values of language models has been pronounced due to the growing need of more authentic, human-aligned responses. However, existing benchmarks rely on human or machine annotations that are vulnerable to value-related biases. Furthermore, the tested scenarios often diverge from real-world contexts in which models are commonly used to generate text and express values. To address these issues, we propose the Value Portrait benchmark, a reliable framework for evaluating LLMs' value orientations with two key characteristics. First, the benchmark consists of items that capture real-life user-LLM interactions, enhancing the relevance of assessment results to real-world LLM usage. Second, each item is rated by human subjects based on its similarity to their own thoughts, and correlations between these ratings and the subjects' actual value scores are derived. This psychometrically validated approach ensures that items strongly correlated with specific values serve as reliable items for assessing those values. Through evaluating 44 LLMs with our benchmark, we find that these models prioritize Benevolence, Security, and Self-Direction values while placing less emphasis on Tradition, Power, and Achievement values. Also, our analysis reveals biases in how LLMs perceive various demographic groups, deviating from real human data.

Donate or Create? Comparing Data Collection Strategies for Emotion-labeled Multimodal Social Media Posts

Christopher Bagdon, Aidan Combs, Carina Silberer, Roman Klinger

Accurate modeling of subjective phenomena such as emotion expression requires data annotated with authors' intentions. Commonly such data is collected by asking study participants to donate and label genuine content produced in the real world, or create content fitting particu- lar labels during the study. Asking participants to create content is often simpler to implement and presents fewer risks to participant privacy than data donation. However, it is unclear if and how study-created content may differ from genuine content, and how differences may impact models. We collect study-created and genuine multimodal social media posts labeled for emotion and compare them on several dimen- sions, including model performance. We find that compared to genuine posts, study-created posts are longer, rely more on their text and less on their images for emotion expression, and focus more on emotion-prototypical events. The samples of participants willing to donate versus create posts are demographically different. Study-created data is valuable to train models that generalize well to genuine data, but realistic effectiveness estimates require genuine data.

Chinese SimpleQA: A Chinese Factuality Evaluation for Large Language Models

Yancheng He, Shilong Li, Jiaheng Liu, Yingshui Tan, Weixun Wang, Hui Huang, Xingyuan Bu, Hangyu Guo, Chengwei Hu, Boren Zheng, Zhuoran Lin, Dekai Sun, Zhicheng Zheng, Wenbo Su, Bo Zheng

New LLM benchmarks are important to align with the rapid development of Large Language Models (LLMs). In this work, we present Chinese SimpleQA, the first comprehensive Chinese benchmark to evaluate the factuality ability of LLMs to answer short questions, and Chinese SimpleQA mainly has five properties (i.e., Chinese, Diverse, High-quality, Static, Easy-to-evaluate). Specifically, first, we focus on the Chinese language over 6 major topics with 99 diverse subtopics. Second, we conduct a comprehensive quality control process to achieve high-quality questions and answers, where the reference answers are static and cannot be changed over time. Third, following SimpleQA, the questions and answers are very short, and the grading process is easy-to-evaluate. Based on Chinese SimpleQA, we perform a comprehensive evaluation of the factuality abilities of existing LLMs. Finally, we hope that Chinese SimpleQA could guide the developers to better understand the Chinese factuality abilities of their models and facilitate the growth of LLMs.

From Tools to Teammates: Evaluating LLMs in Multi-Session Coding Interactions

Nathanaël Carraz Rakotonirina, Mohammed Hamdy, Jon Ander Campos, Lucas Weber, Alberto Testoni, Marzieh Fadaee, Sandro Pezzelle, Marco Del Tredici

Large Language Models (LLMs) are increasingly used in working environments for a wide range of tasks, excelling at solving individual problems in isolation. However, are they also able to effectively collaborate over long-term interactions? To investigate this, we introduce MemoryCode, a synthetic multi-session dataset designed to test LLMs' ability to track and execute simple coding instructions amid irrelevant information, simulating a realistic setting. While all the models we tested handle isolated instructions well, even the performance of state-of-the-art models like GPT-4o deteriorates when instructions are spread across sessions. Our analysis suggests this is due to their failure to retrieve and integrate information over long interaction chains. Our results highlight a fundamental limitation of current LLMs, restricting their ability to collaborate effectively in long interactions.

Text-to-ES Bench: A Comprehensive Benchmark for Converting Natural Language to Elasticsearch Query

DonggeXue, Zhili Pu, Zhentao Xia, Hongli Sun, Ruihui Hou, Guangya Yu, Yupian Lin, Yongqi Fan, Jingping Liu, Tong Ruan

Elasticsearch (ES) is a distributed RESTful search engine optimized for large-scale and long-text search scenarios. Recent research on text-to-Query has explored using large language models (LLMs) to convert user query intent to executable code, making it an increasingly popular research topic. To our knowledge, we are the first to introduce the novel semantic parsing task text-to-ES. To bridge the gap between LLM and ES, in detail, we leverage LLMs and employ domain experts to generate ES query bodies, which are Domain-Specific Language (DSL), along with the corresponding post-processing code to support multi-index ES queries. Consequently, we propose the text-to-ES benchmark that consists of two datasets: Large Elasticsearch Dataset (LED), containing 26,207 text-ES pairs derived from a 224.9GB schema-free database, and ElasticSearch (BirdES)with 10,926 pairs sourced from the Bird dataset on a 33.4GB schema-fixed database. Compared with fourteen advanced LLMs and six code-based LLMs, the model we trained outperformed DeepSeek-R1 by 15.64% on the LED dataset, setting a new state-of-the-art, and achieved 78% of DeepSeek-R1's performance on the BirdES dataset. Additionally, we provide in-depth experimental analyses and suggest future research directions for this task. Our datasets are available at https://huggingface.co/datasets/Barry1915/Text-to-ES.

Evaluating Personalized Tool-Augmented LLMs from the Perspectives of Personalization and Proactivity

Yupu Hao, Pengfei Cao, Zhuoran Jin, Huanxuan Liao, Yubo Chen, Kang Liu, Jun Zhao

Personalized tool utilization is essential for aligning large language models (LLMs) with user preference in interaction scenarios with various tools. However, most of the current benchmarks primarily focus on either personalization of text generation or direct tool-utilizing, without considering both. In this work, we introduce a novel benchmark \textbf{ETAPP} for evaluating personalized tool invocation, establishing a sandbox environment, and a comprehensive dataset of 800 testing cases covering diverse user profiles. To improve the accuracy of our evaluation, we propose a key-point-based LLM evaluation method, mitigating biases in the LLM-as-a-judge system by manually annotating key points for each test case and providing them to LLM as the reference. Additionally, we evaluate the excellent LLMs and provide an in-depth analysis. Furthermore, we investigate the impact of different tool-invoking strategies on LLMs' personalization performance and the effects of fine-tuning in our task. The effectiveness of our preference-setting and key-point-based evaluation method is also validated. Our findings offer insights into improving personalized LLM agents. Our code is available at https://github.com/hypasd-art/ETAPP.

IndicSynth: A Large-Scale Multilingual Synthetic Speech Dataset for Low-Resource Indian Languages

Divya V Sharma, Vijval Ekbote, Anubha Gupta

Recent advances in synthetic speech generation technology have facilitated the generation of high-quality synthetic (fake) speech that emulates human voices. These technologies pose a threat of misuse for identity theft and the spread of misinformation. Consequently, the misuse of such powerful technologies necessitates the development of robust and generalizable audio deepfake detection (ADD) and anti-spoofing models. However, such models are often linguistically biased. Consequently, the models trained on datasets in one language exhibit a low accuracy when evaluated on out-of-domain languages. Such biases reduce the usability of these models and highlight the urgent need for multilingual synthetic speech datasets for bias mitigation research. However, most available datasets are in English or Chinese. The dearth of multilingual synthetic datasets hinders multilingual ADD and anti-spoofing research. Furthermore, the problem intensifies in countries with rich linguistic diversity, such as India. Therefore, we introduce IndicSynth, which contains 4,000 hours of synthetic speech from 989 target speakers, including 456 females and 533 males for 12 low-resourced Indian languages. The dataset includes rich metadata covering gender details and target speaker identifiers. Experimental results demonstrate that IndicSynth is a valuable contribution to multilingual ADD and anti-spoofing research. The dataset can be accessed from https://github.com/vdivyas/IndicSynth.

Condor: Enhance LLM Alignment with Knowledge-Driven Data Synthesis and Refinement

Maosongcao, Taolin Zhang, Mo Li, Chuyu Zhang, Yunxin Liu, Conghui He, Haodong Duan, Songyang Zhang, Kai Chen

The quality of Supervised Fine-Tuning (SFT) data plays a critical role in enhancing the conversational capabilities of Large Language Models (LLMs). However, the availability of high-quality human-annotated SFT data has become a significant bottleneck for LLMs, necessitating a greater reliance on synthetic training data. In this work, we introduce Condor, a two-stage synthetic data generation framework that incorporates World Knowledge Trees and Self-Reflection Refinement to produce high-quality SFT data at scale. Our experimental results demonstrate that a base model fine-tuned on only 20K Condor-generated samples achieves superior performance compared to instruct model trained with RLHF. The additional refinement stage in Condor further enables iterative self-improvement for LLMs at various scales (up to 72B), validating the effectiveness of our approach. Furthermore, our investigation into the scaling of synthetic data in post-training reveals substantial unexplored potential for performance improvements, opening promising avenues for future research.

Silencing Empowerment, Allowing Bigotry: Auditing the Moderation of Hate Speech on Twitch

Prarabdh Shukla, Wei Yin Chong, Yash Patel, Brennan Schaffner, Danish Pruthi, Arjun Bhagoji

To meet the demands of content moderation, online platforms have resorted to automated systems. Newer forms of real-time engagement (\textit{e.g.}, users commenting on live streams) on platforms like Twitch exert additional pressures on the latency expected of such moderation systems. Despite their prevalence, relatively little is known about the effectiveness of these systems. In this paper, we conduct an audit of Twitch's automated moderation tool (\texttt{AutoMod}) to investigate its effectiveness in flagging hateful content. For our audit, we create streaming accounts to act as siloed test beds, and interface with the live chat using Twitch's APIs to send over 107,000 comments collated from 4 datasets. We measure \texttt{AutoMod}'s accuracy in flagging blatantly hateful content containing misogyny, racism, ableism and homophobia. Our experiments reveal that a large fraction of hateful messages, up to 94% on some datasets, \text{\textit{bypass moderation}}. Contextual addition of slurs to these messages results in 100% removal, revealing \texttt{AutoMod}'s reliance on slurs as a hate signal. We also find that contrary to Twitch's community guidelines, \texttt{AutoMod} blocks up to 89.5% of benign examples that use sensitive words in pedagogical or empowering contexts. Overall, our audit points to large gaps in \texttt{AutoMod}'s capabilities and underscores the importance for such systems to understand context effectively.

CRUXEVAL-X: A Benchmark for Multilingual Code Reasoning, Understanding and Execution

Ruiyang Xu, Jialun Cao, Yaojie Lu, Ming Wen, Hongyu Lin, Xianpei Han, Ben He, Shing-Chi Cheung, Le Sun

Code benchmarks such as HumanEval are widely adopted to evaluate Large Language Models' (LLMs) coding capabilities. However, there is an unignorable programming language bias in existing code benchmarks -- over 95% code generation benchmarks are dominated by Python, leaving the LLMs' capabilities in other programming languages such as Java and C/C++ unknown. Moreover, coding task bias is also crucial. Most benchmarks focus on code generation capability, while benchmarks for code reasoning (given input, reasoning output; and given output, reasoning input), an essential coding capability, are insufficient. Yet, constructing multi-lingual benchmarks can be expensive and labor-intensive, and codes in contest websites such as Leetcode suffer from data contamination during training. To fill this gap, we propose CRUXEVAL-X, a multi-lingual code reasoning benchmark that contains 19 programming languages. It comprises at least 600 subjects for each language, along with 19K content-consistent tests in total. In particular, the construction pipeline of CRUXEVAL-X works in a fully automated and test-guided manner, which iteratively generates and repairs based on execution feedback. Also, to cross language barriers (e.g., dynamic/static type systems in Python/C++), we formulated various transition rules between language pairs to facilitate translation. Our intensive evaluation of 24 representative LLMs reveals the correlation between language pairs. For example, TypeScript and JavaScript show a significant positive correlation, while Racket has less correlation with other languages. More interestingly, even a model trained solely on Python can achieve at most 34.4% Pass@1 in other languages, revealing the cross-language generalization of LLMs.

Evaluating LLMs for Portuguese Sentence Simplification with Linguistic Insights

ARTHUR MARIANO ROCHA DE AZEVEDO SCALERCIO, Elvis A. de Souza, Maria José Bocorny Finatto, Aline Paes

Sentence simplification (SS) focuses on adapting sentences to enhance their readability and accessibility. While large language models (LLMs) match task-specific baselines in English SS, their performance in Portuguese remains underexplored. This paper presents a comprehensive performance comparison of 26 state-of-the-art LLMs in Portuguese SS, alongside two simplification models trained explicitly for this task and language. They are evaluated under a one-shot setting across scientific, news, and government datasets. We benchmark the models with our newly introduced Gov-Lang-BR corpus (1,703 complex-simple sentence pairs from Brazilian government agencies) and two established datasets: PorSimplesSent and Museum-PT. Our investigation takes advantage of both automatic metrics and large-scale linguistic analysis to examine the transformations achieved by the LLMs. Furthermore, a qualitative assessment of selected generated outputs provides deeper insights into simplification quality. Our findings reveal that while open-source LLMs have achieved impressive results, closed-source LLMs continue to outperform them in Portuguese SS.

PRMBench: A Fine-grained and Challenging Benchmark for Process-Level Reward Models

Mingyang Song, Zhaochen Su, Xiaoye Qu, Jiawei Zhou, Yu Cheng

Process-level Reward Models (PRMs) are crucial for complex reasoning and decision-making tasks, where each intermediate step plays an important role in the reasoning process. Since language models are prone to various types of errors during the reasoning process, PRMs are required to possess nuanced capabilities for detecting various implicit error types in real-world scenarios. However, current benchmarks primarily focus on step correctness, failing to evaluate PRMs' performance systematically. To address this gap, we introduce PRMBench, a process-level benchmark specifically designed to assess the fine-grained error detection capabilities of PRMs. PRMBench comprises 6,216 carefully designed problems and 83,456 step-level labels, evaluating models across multiple dimensions, including \textit{simplicity}, \textit{soundness}, and \textit{sensitivity}. In our experiments on 25 models, spanning both open-source PRMs and closed-source large language models prompted as critic models, we uncover significant weaknesses in current PRMs. These findings underscore the challenges inherent in process-level evaluation and highlight key directions for future research, establishing PRMBench as a robust testbed for advancing research on PRM evaluation and development.

On Many-Shot In-Context Learning for Long-Context Evaluation

Kaijian Zou, Muhammad Khalifa, Lu Wang

Many-shot in-context learning (ICL) has emerged as a unique setup to both utilize and test the ability of large language models to handle long context. This paper delves into long-context language model (LCLM) evaluation through many-shot ICL. We first ask: what types of ICL tasks benefit from additional demonstrations, and how effective are they in evaluating LCLMs? We find that classification and summarization tasks show performance improvements with additional demonstrations, while translation and reasoning tasks do not exhibit clear trends. Next, we investigate the extent to which different tasks necessitate retrieval versus global context understanding. We develop metrics to categorize ICL tasks into two groups: (i) similar-sample learning (SSL): tasks where retrieval of the most similar examples is sufficient for good performance, and (ii) all-sample learning (ASL): tasks that necessitate a deeper comprehension of all examples in the prompt. Lastly, we introduce a new many-shot ICL benchmark built on existing ICL tasks, MANYICLBENCH, to characterize model's ability on both fronts and benchmark 12 LCLMs using MANYICLBENCH. We find that while state-of-the-art models demonstrate good performance up to 64k tokens in SSL tasks, many models experience significant performance drops at only 16k tokens in ASL tasks.

Benchmarking Long-Context Language Models on Long Code Understanding

Jia Li, Xuyuan Guo, Lei Li, Kechi Zhang, Ge Li, Jia Li, Zhengwei Tao, Fang Liu, Chongyang Tao, Yuqi Zhu, Zhi Jin

Current advanced long-context language models offer great potential for real-world software engineering applications. However, progress in this critical domain remains hampered by a fundamental limitation: the absence of a rigorous evaluation framework for long code understanding. To gap this obstacle, we propose a long code understanding benchmark LongCodeU from four aspects (8 tasks) to evaluate LCLMs' long code understanding ability required for practical applications, including code unit perception, intra-code unit understanding, inter-code unit relation understanding, and long code documentation understanding. We evaluate 9 popular LCLMs on LongCodeU (i.e., 6 general models and 3 code models). Our experimental results reveal key limitations in current LCLMs' capabilities for long code understanding. Particularly, the performance of LCLMs drops dramatically when the long code length is greater than 32K, falling far short of their claimed 128K to 1M context windows. In the four aspects, inter-code unit relation understanding is the most challenging for LCLMs. Our study provides valuable insights for optimizing LCLMs and driving advancements in software engineering.

Memorizing is Not Enough: Deep Knowledge Injection Through Reasoning

Ruoxi Xu, Yunjie Ji, Boxi Cao, Yaojie Lu, Hongyu Lin, Xianpei Han, Ben He, Yingfei Sun, Xiangang Li, Le Sun

Although large language models (LLMs) excel in knowledge recall and reasoning, their static nature leads to outdated information as the real world evolves or when adapting to domain-specific knowledge, highlighting the need for effective knowledge injection. However, current research on knowledge injection remains superficial, mainly focusing on knowledge memorization and retrieval. This paper proposes a four-tier knowledge injection framework that systematically defines the levels of knowledge injection: memorization, retrieval, reasoning, and association. Based on this framework, we introduce DeepKnowledge, a synthetic experimental testbed designed for fine-grained evaluation of the depth of knowledge injection across three knowledge types (novel, incremental, and updated). We then explore various knowledge injection scenarios and evaluate the depth of knowledge injection for each scenario on the benchmark. Experimental results reveal key factors to reach each level of knowledge injection for LLMs and establish a mapping between the levels of knowledge injection and the corresponding suitable injection methods, aiming to provide a comprehensive approach for efficient knowledge injection across various levels. The code is available at https://github.com/icip-cas/Knowledge-Learning-Toolkits.

ConLoan: A Contrastive Multilingual Dataset for Evaluating Loanwords

Sina Ahmadi, Micha David Hess, Elena Álvarez-Mellado, Alessia Battisti, Cui Ding, Anne Göhring, Yingqiang Gao, Zifan Jiang, Andrianos Michail, Peshmerge Morad, Joel Niklaus, Maria Christina Panagiotopoulou, Stefano Perrella, Juri Opitz, Anastassia Shaitarova, Rico Sennrich

Lexical borrowing, the adoption of words from one language into another, is a ubiquitous linguistic phenomenon influenced by geopolitical, societal, and technological factors. This paper introduces ConLoan--a novel contrastive dataset comprising sentences with and without loanwords across 10 languages. Through systematic evaluation using this dataset, we investigate how state-of-the-art machine translation and language models process loanwords compared to their native alternatives. Our experiments reveal that these systems show systematic preferences for loanwords over native terms and exhibit varying performance across languages. These findings provide valuable insights for developing more linguistically robust NLP systems.

Benchmarking LLMs and LLM-based Agents in Practical Vulnerability Detection for Code Repositories

Alperen Yildiz, Sin G Teo, Yiling Lou, Yebo Feng, Chong Wang, Dinil Mon Divakaran

Large Language Models (LLMs) have shown promise in software vulnerability detection, particularly on function-level benchmarks like Devign and BigVul. However, real-world detection requires interprocedural analysis, as vulnerabilities often emerge through multi-hop function calls rather than isolated functions. While repository-level benchmarks like ReposVul and VulEval introduce interprocedural context, they remain computationally expensive, lack pairwise evaluation of vulnerability fixes, and explore limited context retrieval, limiting their practicality. We introduce JITVul, a JIT vulnerability detection benchmark linking each function to its vulnerability-introducing and fixing commits. Built from 879 CVEs spanning 91 vulnerability types, JITVul enables comprehensive evaluation of detection capabilities. Our results show that ReAct Agents, leveraging thought-action-observation and interprocedural context, perform better than LLMs in distinguishing vulnerable from benign code. While prompting strategies like Chain-of-Thought help LLMs, ReAct Agents require further refinement. Both methods show inconsistencies, either misidentifying vulnerabilities or over-analyzing security guards, indicating significant room for improvement.

Entailed Between the Lines: Incorporating Implication into NLI

Shreya Havaldar, Hamidreza Alvari, John Palowitch, Mohammad Javad Hosseini, Senaka Buthpitiya, Alex Fabrikant

Much of human communication depends on implication, conveying meaning beyond literal words to express a wider range of thoughts, intentions, and feelings. For models to better understand and facilitate human communication, they must be responsive to the text's implicit meaning. We focus on Natural Language Inference (NLI), a core tool for many language tasks, and find that state-of-the-art NLI models and datasets struggle to recognize a range of cases where entailment is implied, rather than explicit from the text. We formalize implied entailment as an extension of the NLI task and introduce the Implied NLI dataset (INLI) to help today's LLMs both recognize a broader variety of implied entailments and to distinguish between implicit and explicit entailment. We show how LLMs fine-tuned on INLI understand implied entailment and can generalize this understanding across datasets and domains.

Towards Building Large Scale Datasets and State-of-the-Art Automatic Speech Translation Systems for 13 Indian Languages

Ashwin Sankar, Sparsh Jain, Nikhil Narasimhan, Devilal Choudhary, Dhairya Suman, Mohammed Safi Ur Rahman Khan, Anoop Kunchukuttan, Mitesh M Khapra, Raj Dabre

Speech translation for Indian languages remains a challenging task due to the scarcity of large-scale, publicly available datasets that capture the linguistic diversity and domain coverage essential for real-world applications. Existing datasets cover a fraction of Indian languages and lack the breadth needed to train robust models that generalize beyond curated benchmarks. To bridge this gap, we introduce BhasaAnuvaad, the largest speech translation dataset for Indian languages, spanning over 44 thousand hours of audio and 17 million aligned text segments across 14 Indian languages and English. Our dataset is built through a threefold methodology: (a) aggregating high-quality existing sources, (b) large-scale web crawling to ensure linguistic and domain diversity, and (c) creating synthetic data to model real-world speech disfluencies. Leveraging BhasaAnuvaad, we train IndicSeamless, a state-of-the-art speech translation model for Indian languages that performs better than existing models. Our experiments demonstrate improvements in the translation quality, setting a new standard for Indian language speech translation. We will release all the code, data and model weights in the open-source, with permissive licenses to promote accessibility and collaboration.

Session 2: Semantics: Lexical and Sentence-Level (Monday 28 July, 11:00-12:30, Hall 4/5)

How Much Do Pretrained Language Models Know About Word Senses?

Simone Teglia, Simone Tedeschi, Roberto Navigli

Word Sense Disambiguation (WSD) is a key task in Natural Language Processing (NLP), involving selecting the correct meaning of a word based on its context. With Pretrained Language Models (PLMs) like BERT and DeBERTa now well established, significant progress has been made in understanding contextual semantics. Nevertheless, how well these models inherently disambiguate word senses remains uncertain. In this work, we evaluate several encoder-only PLMs across two popular inventories (i.e. WordNet and the Oxford Dictionary of English) by analyzing their ability to separate word senses without any task-specific fine-tuning. We compute centroids of word senses and measure similarity to assess performance across different layers. Our results show that DeBERTa-v3 delivers the best performance on the task, with the middle layers (specifically the 7th and 8th layers) achieving the highest accuracy, outperforming the output layer by approximately 15 percentage points. Our experiments also explore the inherent structure of WordNet and ODE sense inventories, highlighting their influence on the overall model behavior and performance. Finally, based on our findings, we develop a small, efficient model for the WSD task that attains robust performance while significantly reducing the carbon footprint.

Entailment-Preserving First-order Logic Representations in Natural Language Entailment

Jinu Lee, Qi Liu, Runzhi Ma, Vincent Han, Ziqi Wang, Heng Ji, Julia Hockenmaier

First-order logic (FOL) is often used to represent logical entailment, but determining natural language (NL) entailment using FOL remains a challenge. To address this, we propose the Entailment-Preserving FOL representations (EPF) task and introduce reference-free evaluation metrics for EPF (Entailment-Preserving Rate (EPR) family). In EPF, one should generate FOL representations from multi-premise NL entailment data (e.g., EntailmentBank) so that the automatic prover's result preserves the entailment labels. Furthermore, we propose a training method specialized for the task, iterative learning-to-rank, which trains an NL-to-FOL translator by using the natural language entailment labels as verifiable rewards. Our method achieves a 1.8–2.7% improvement in EPR and a 17.4–20.6% increase in EPR@16 compared to diverse baselines in three datasets. Further analyses reveal that iterative learning-to-rank effectively suppresses the arbitrariness of FOL representation by reducing the diversity of predicate signatures, and maintains strong performance across diverse inference types and out-of-domain data.

Uncertainty in Causality: A New Frontier

Shaobo Cui, Luca Mouchel, Boi Faltings

Understanding uncertainty in causality is vital in various domains, including core NLP tasks like event causality extraction, commonsense reasoning, and counterfactual text generation. However, existing literature lacks a comprehensive examination of this area. This survey aims to fill this gap by thoroughly reviewing uncertainty in causality. We first introduce a novel trichotomy, categorizing causal uncertainty into aleatoric (inherent randomness in causal data), epistemic (causal model limitations), and ontological (existence of causal links) uncertainty. We then survey methods for quantifying uncertainty in causal analysis and highlight the complementary relationship between causal uncertainty and causal strength. Furthermore, we examine the challenges that large language models (LLMs) face in handling causal uncertainty, such as hallucinations and inconsistencies, and propose key traits for an optimal causal LLM. Our paper reviews current approaches and outlines future research directions, aiming to serve as a practical guide for researchers and practitioners in this emerging field.

Quantifying Lexical Semantic Shift via Unbalanced Optimal Transport

Ryo Kishino, Hiroaki Yamagiwa, Ryo Nagata, Sho Yokoi, Hidetoshi Shimodaira

Lexical semantic change detection aims to identify shifts in word meanings over time. While existing methods using embeddings from a diachronic corpus pair estimate the degree of change for target words, they offer limited insight into changes at the level of individual usage instances. To address this, we apply Unbalanced Optimal Transport (UOT) to sets of contextualized word embeddings, capturing semantic change through the excess and deficit in the alignment between usage instances. In particular, we propose Sense Usage Shift (SUS), a measure that quantifies changes in the usage frequency of a word sense at each usage instance. By leveraging SUS, we demonstrate that several challenges in semantic change detection can be addressed in a unified manner, including quantifying instance-level semantic change and word-level tasks such as measuring the magnitude of semantic change and the broadening or narrowing of meaning.

Faithful and Robust LLM-Driven Theorem Proving for NLI Explanations

Xin Quan, Marco Valentino, Louise A. Dennis, Andre Freitas

Natural language explanations play a fundamental role in Natural Language Inference (NLI) by revealing how premises logically entail hypotheses. Recent work has shown that the interaction of large language models (LLMs) with theorem provers (TPs) can help verify and improve the validity of NLI explanations. However, TPs require translating natural language into machine-verifiable formal representations, a process that introduces the risk of semantic information loss and unfaithful interpretation, an issue compounded by LLMs' challenges in capturing critical logical structures with sufficient precision. Moreover, LLMs are still limited in their capacity for rigorous and robust proof construction within formal verification frameworks. To mitigate issues related to faithfulness and robustness, this paper investigates strategies to (1) alleviate semantic loss during autoformalisation, (2) efficiently identify and correct syntactic errors in logical representations, (3) explicitly use logical expressions to guide LLMs in generating structured proof sketches, and (4) increase LLMs' capacity of interpreting TP's feedback for iterative refinement. Our empirical results on e-SNLI, QASC and WorldTree using different LLMs demonstrate that the proposed strategies yield significant improvements in autoformalisation (+18.46%, +34.2%, +39.77%) and explanation refinement (+29.5%, +51.5%, +41.25%) over the state-of-the-art model. Moreover, we show that specific interventions on the hybrid LLM-TP architecture can substantially improve efficiency, drastically reducing the number of iterations required for successful verification.

OASIS: Order-Augmented Strategy for Improved Code Search

GAO Zuchen, Zizheng Zhan, Xianming LI, Erxin Yu, Haotian Zhang, chenbin, Yuqun Zhang, Jing Li

Code embeddings capture the semantic representations of code and are crucial for various code-related large language model (LLM) applications, such as code search. Previous training primarily relies on optimizing the InfoNCE loss by comparing positive natural language (NL)-code pairs with in-batch negatives. However, due to the sparse nature of code contexts, training solely by comparing the major differences between positive and negative pairs may fail to capture deeper semantic nuances. To address this issue, we propose a novel order-augmented strategy for improved code search (OASIS). It leverages order-based similarity labels to train models to capture subtle differences in similarity among negative pairs. Extensive benchmark evaluations demonstrate that our OASIS model significantly outperforms previous state-of-the-art models focusing solely on major positive-negative differences. It underscores the value of exploiting subtle differences among negative pairs with order labels for effective code embedding training.

Partial Colexifications Improve Concept Embeddings

Arne Rubehn, Johann-Mattis List

While the embedding of words has revolutionized the field of Natural Language Processing, the embedding of concepts has received much less attention so far. A dense and meaningful representation of concepts, however, could prove useful for several tasks in computational linguistics, especially those involving cross-linguistic data or sparse data from low resource languages. First methods that have been proposed so far embed concepts from automatically constructed colexification networks. While these approaches depart from automatically inferred polysemies, attested across a larger number of languages, they are restricted to the word level, ignoring lexical relations that would only hold for parts of the words in a given language. Building on recently introduced methods for the inference of partial colexifications, we show how they can be used to improve concept embeddings in meaningful ways. The learned embeddings are evaluated against lexical similarity ratings, recorded instances of semantic shift, and word association data. We show that in all evaluation tasks, the inclusion of partial colexifications lead to improved concept representations and better results. Our results further show that the learned embeddings are able to capture and represent different semantic relationships between concepts.

On the Relation Between Fine-Tuning, Topological Properties, and Task Performance in Sense-Enhanced Embeddings

Deniz Ekin Yavas, Timothée Bernard, Benoit Crabbé, Laura Kallmeyer

Topological properties of embeddings, such as isotropy and uniformity, are closely linked to their expressiveness, and improving these properties enhances the embeddings' ability to capture nuanced semantic distinctions. However, fine-tuning can reduce the expressiveness of the embeddings of language models. This study investigates the relation between fine-tuning, topology of the embedding space, and task performance in the context of sense knowledge enhancement, focusing on identifying the topological properties that contribute to the success of sense-enhanced embeddings. We experiment with two fine-tuning methods: *Supervised Contrastive Learning (SCL)* and *Supervised Predictive Learning (SPL)*. Our results show that SPL, the most standard approach, exhibits varying effectiveness depending on the language model and is inconsistent in producing successful sense-enhanced embeddings. In contrast, SCL achieves this consistently. Furthermore, while the embeddings with only increased *sense-alignment* show reduced task performance, those that also exhibit high *isotropy* and balance *uniformity* with *sense-alignment* achieve the best results. Additionally, our findings indicate that supervised and unsupervised tasks benefit from these topological properties to varying degrees.

Causal Graph based Event Reasoning using Semantic Relation Experts

Mahnaz Koupaee, Xueying Bai, Mudan Chen, Greg Durrett, Nathanael Chambers, Niranjan Balasubramanian

Understanding how events in a scenario causally connect with each other is important for effectively modeling and reasoning about events. But event reasoning remains a difficult challenge, and despite recent advances, Large Language Models (LLMs) still struggle to accurately identify causal connections between events. This struggle leads to poor performance on deeper reasoning tasks like event forecasting and timeline understanding. To address this challenge, we investigate the generation of causal event graphs (e.g., A enables B) as a parallel mechanism to help LLMs explicitly represent causality during inference. This paper evaluates both how to generate correct graphs as well as how graphs can assist reasoning. We propose a collaborative approach to causal graph generation where we use LLMs to simulate experts that focus on specific semantic relations. The experts engage in multiple rounds of discussions which are then consolidated by a final expert. Then, to demonstrate the utility of causal graphs, we use them on multiple downstream applications, and also introduce a new explainable event prediction task that requires a causal chain of events in the explanation. These explanations are more informative and coherent than baseline generations. Finally, our overall approach not finetuned on any downstream task, achieves competitive results with state-of-the-art models on both forecasting and next event prediction tasks.

CoAM: Corpus of All-Type Multiword Expressions

Yusuke Ide, Joshua Tanner, Adam Nohejl, Jacob Hoffman, Justin Vasselli, Hidetaka Kamigaito, Taro Watanabe

Multiword expressions (MWEs) refer to idiomatic sequences of multiple words. MWE identification, i.e., detecting MWEs in text, can play a key role in downstream tasks such as machine translation, but existing datasets for the task are inconsistently annotated, limited to a single type of MWE, or limited in size. To enable reliable and comprehensive evaluation, we created CoAM: Corpus of All-Type Multiword Expressions, a dataset of 1.3K sentences constructed through a multi-step process to enhance data quality consisting of human annotation, human review, and automated consistency checking. Additionally, for the first time in a dataset of MWE identification, CoAM's MWEs are tagged with MWE types, such as Noun and Verb, enabling fine-grained error analysis. Annotations for CoAM were collected using a new interface created with our interface generator, which allows easy and flexible annotation of MWEs in any form. Through experiments using CoAM, we find that a fine-tuned large language model outperforms MWEasWSD, which achieved the state-of-the-art performance on the DiMSUM dataset. Furthermore, analysis using our MWE type tagged data reveals that Verb MWEs are easier than Noun MWEs to identify across approaches.

Sticking to the Mean: Detecting Sticky Tokens in Text Embedding Models

Kexin Chen, Dongxia Wang, Yi Liu, Haonan Zhang, Wenhai Wang

Despite the widespread use of Transformer-based text embedding models in NLP tasks, surprising ``sticky tokens'' can undermine the reliability of embeddings. These tokens, when repeatedly inserted into sentences, pull sentence similarity toward a certain value, disrupting the normal distribution of embedding distances and degrading downstream performance. In this paper, we systematically investigate such anomalous tokens, formally defining them and introducing an efficient detection method, Sticky Token Detector (STD), based on sentence and token filtering. Applying STD to 40 checkpoints across 14 model families, we discover a total of 868 sticky tokens. Our analysis reveals that these tokens often originate from special or unused entries in the vocabulary, as well as fragmented subwords from multilingual corpora. Notably, their presence does not strictly correlate with model size or vocabulary size. We further evaluate how sticky tokens affect downstream tasks like clustering and retrieval, observing significant performance drops of up to 50\%. Through attention-layer analysis, we show that sticky tokens disproportionately dominate the model’s internal representations, raising concerns about tokenization robustness. Our findings show the need for better tokenization strategies and model design to mitigate the impact of sticky tokens in future text embedding applications.

Negative Matters: Multi-Granularity Hard-Negative Synthesis and Anchor-Token-Aware Pooling for Enhanced Text Embeddings

Tengyu Pan, Zhichao Duan, Zhenyu Li, Bowen Dong, Ning Liu, Xiuxing Li, Jianyong Wang

Text embedding models are essential for various natural language processing tasks, enabling the effective encoding of semantic information into dense vector representations. These models are typically optimized using triplets of (query, positive, negative) data pairs for contrastive learning, where the negative samples play a critical role in enhancing the model's ability to discern subtle semantic distinctions. In this work, we introduce a **M**ulti-**G**ranularity **H**ard-negative (MGH) synthesis framework that leverages large language models (LLMs) to generate diverse negative samples with varying levels of similarity with the query. This approach facilitates a coarse-to-fine curriculum learning strategy during supervised training, allowing the embedding model to progressively learn more nuanced semantic representations. Meanwhile, we propose an **A**nchor **T**oken **A**ware (ATA) pooling method that assigns higher weights to anchor tokens based on aggregation patterns observed in LLMs, improving text embedding accuracy without increasing model complexity. Comprehensive experiments on the MTEB benchmark demonstrate that our methods achieve state-of-the-art performance, surpassing existing synthesis strategies both with synthetic data and when combined with public retrieval datasets.

Session 2: Sentiment Analysis, Stylistic Analysis, and Argument Mining (Monday 28 July, 11:00-12:30, Hall 4/5)

FACT-AUDIT: An Adaptive Multi-Agent Framework for Dynamic Fact-Checking Evaluation of Large Language Models

Hongzhan Lin, Yang Deng, Yuxuan Gu, Wenxuan Zhang, Jing Ma, See-Kiong Ng, Tat-Seng Chua

Large Language Models (LLMs) have significantly advanced the fact-checking studies. However, existing automated fact-checking evaluation methods rely on static datasets and classification metrics, which fail to automatically evaluate the justification production and uncover the nuanced limitations of LLMs in fact-checking. In this work, we introduce FACT-AUDIT, an agent-driven framework that adaptively and dynamically assesses LLMs' fact-checking capabilities. Leveraging importance sampling principles and multi-agent collaboration, FACT-AUDIT generates adaptive and scalable datasets, performs iterative model-centric evaluations, and updates assessments based on model-specific responses. By incorporating justification production alongside verdict prediction, this framework provides a comprehensive and evolving audit of LLMs' factual reasoning capabilities, to investigate their trustworthiness. Extensive experiments demonstrate that FACT-AUDIT effectively differentiates among state-of-the-art LLMs, providing valuable insights into model strengths and limitations in model-centric fact-checking analysis.

Enhancing Hyperbole and Metaphor Detection with Their Bidirectional Dynamic Interaction and Emotion Knowledge

Li Zheng, Sihang Wang, Hao Fei, Zuquan Peng, Fei Li, Jianming Fu, Chong Teng, Donghong Ji

Text-based hyperbole and metaphor detection are of great significance for natural language processing (NLP) tasks. However, due to their semantic obscurity and expressive diversity, it is rather challenging to identify them. Existing methods mostly focus on superficial text features, ignoring the associations of hyperbole and metaphor as well as the effect of implicit emotion on perceiving these rhetorical devices. To implement these hypotheses, we propose an emotion-guided hyperbole and metaphor detection framework based on bidirectional dynamic interaction (EmoBi). Firstly, the emotion analysis module deeply mines the emotion connotations behind hyperbole and metaphor. Next, the emotion-based domain mapping module identifies the target and source domains to gain a deeper understanding of the implicit meanings of hyperbole and metaphor. Finally, the bidirectional dynamic interaction module enables the mutual promotion between hyperbole and metaphor. Meanwhile, a verification mechanism is designed to ensure detection accuracy and reliability. Experiments show that EmoBi outperforms all baseline methods on four datasets. Specifically, compared to the current SoTA, the F1 score increased by 28.1\% for hyperbole detection on the TroFi dataset and 23.1\% for metaphor detection on the HYPO-L dataset. These results, underpinned by in-depth analyses, underscore the effectiveness and potential of our approach for advancing hyperbole and metaphor detection.

LACA: Improving Cross-lingual Aspect-Based Sentiment Analysis with LLM Data Augmentation

Jakub Šmíd, Pavel Priban, Pavel Kral

Cross-lingual aspect-based sentiment analysis (ABSA) involves detailed sentiment analysis in a target language by transferring knowledge from a source language with available annotated data. Most existing methods depend heavily on often unreliable translation tools to bridge the language gap. In this paper, we propose a new approach that leverages a large language model (LLM) to generate high-quality pseudo-labelled data in the target language without the need for translation tools. First, the framework trains an ABSA model to obtain predictions for unlabelled target language data. Next, LLM is prompted to generate natural sentences that better represent these noisy predictions than the original text. The ABSA model is then further fine-tuned on the resulting pseudo-labelled dataset. We demonstrate the effectiveness of this method across six languages and five backbone models, surpassing previous state-of-the-art translation-based approaches. The proposed framework also supports generative models, and we show that fine-tuned LLMs outperform smaller multilingual models.

Multimodal Transformers are Hierarchical Modal-wise Heterogeneous Graphs

Yijie Jin, Junjie Peng, Xuanchao Lin, Haochen Yuan, Lan Wang, Cangzhi Zheng

Multimodal Sentiment Analysis (MSA) is a rapidly developing field that integrates multimodal information to recognize sentiments, and existing models have made significant progress in this area. The central challenge in MSA is multimodal fusion, which is predominantly addressed by Multimodal Transformers (MulTs). Although act as the paradigm, MulTs suffer from efficiency concerns. In this work, from the perspective of efficiency optimization, we propose and prove that MulTs are hierarchical modal-wise heterogeneous graphs (HMHGs), and we introduce the graph-structured representation pattern of MulTs. Based on this pattern, we propose an Interlaced Mask (IM) mechanism to design the Graph-Structured and Interlaced-Masked Multimodal Transformer (GsiT). It is formally equivalent to MulTs which achieves an efficient weight-sharing mechanism without information disorder through IM, enabling All-Modal-In-One fusion with only 1/3 of the parameters of pure MulTs. A kernel called Decomposition is implemented to ensure avoiding additional computational overhead. Moreover, it achieves significantly higher performance than traditional MulTs. To further validate the effectiveness of GsiT itself and the HMHG concept, we integrate them into multiple state-of-the-art models and demonstrate notable performance improvements and parameter reduction on widely used MSA datasets. Experimental results also demonstrate its effectiveness on other multimodal tasks. The code is available in https://github.com/drewjin/GsiT.git.

Context-Aware Sentiment Forecasting via LLM-based Multi-Perspective Role-Playing Agents

Fanhang Man, Huandong Wang, Jianjie Fang, Zhaoyi Deng, Baining Zhao, Xinlei Chen, Yong Li

User sentiment on social media reveals underlying social trends, crises, and needs. Researchers have analyzed users' past messages to track the evolution of sentiments and reconstruct sentiment dynamics. However, predicting the imminent sentiment response of users to ongoing events remains understudied. In this paper, we address the problem of sentiment forecasting on social media to predict users' future sentiment based on event developments. We extract sentiment-related features to enhance modeling and propose a multi-perspective role-playing framework to simulate human response processes. Our preliminary results show significant improvements in sentiment forecasting at both microscopic and macroscopic levels.

Multi-level Association Refinement Network for Dialogue Aspect-based Sentiment Quadruple Analysis

Zeliang Tong, Wei Wei, Xiaoye Qu, Rikui Huang, Zhixin Chen, Xingyu Yan

Dialogue Aspect-based Sentiment Quadruple (DiaASQ) analysis aims to identify all quadruples (i.e., target, aspect, opinion, sentiment) from the dialogue. This task is challenging as different elements within a quadruple may manifest in different utterances, requiring precise handling of associations at both the utterance and word levels. However, most existing methods tackling it predominantly leverage predefined dialogue structure (e.g., reply) and word semantics, resulting in a surficial understanding of the deep sentiment association between utterances and words. In this paper, we propose a novel Multi-level Association Refinement Network (MARN) designed to achieve more accurate and comprehensive sentiment associations between utterances and words. Specifically, for utterances, we dynamically capture their associations with enriched semantic features through a holistic understanding of the dialogue, aligning them more closely with sentiment associations within elements in quadruples. For words, we develop a novel cross-utterance syntax parser (CU-Parser) that fully exploits syntactic information to enhance the association between word pairs within and across utterances. Moreover, to address the scarcity of labeled data in DiaASQ, we further introduce a multi-view data augmentation strategy to enhance the performance of MARN under low-resource conditions. Experimental results demonstrate that MARN achieves state-of-the-art performance and maintains robustness even under low-resource conditions.

Learning First-Order Logic Rules for Argumentation Mining

Yang Sun, Guanrong Chen, Hamid Alinejad-Rokny, Jianzhu Bao, Yuqi Huang, Bin Liang, Kam-Fai Wong, Min Yang, Ruifeng Xu

Argumentation Mining (AM) aims to extract argumentative structures from texts by identifying argumentation components (ACs) and their argumentative relations (ARs). While previous works focus on representation learning to encode ACs and AC pairs, they fail to explicitly model the underlying reasoning patterns of AM, resulting in limited interpretability. This paper proposes a novel \underline{F}irst-\underline{O}rder \underline{L}ogic reasoning framework for \underline{AM} (FOL-AM), designed to explicitly capture logical reasoning paths within argumentative texts. By interpreting multiple AM subtasks as a unified relation query task modeled using FOL rules, FOL-AM facilitates multi-hop relational reasoning and enhances interpretability. The framework supports two flexible implementations: a fine-tuned approach to leverage task-specific learning, and a prompt-based method utilizing large language models to harness their generalization capabilities. Extensive experiments on two AM benchmarks demonstrate that FOL-AM outperforms strong baselines while significantly improving explainability.

Controllable Style Arithmetic with Language Models

Weiqi Wang, Wengang Zhou, Zongmeng Zhang, Jie Zhao, Houqiang Li

Language models have shown remarkable capabilities in text generation, but precisely controlling their linguistic style remains challenging. Existing methods either lack fine-grained control, require extensive computation, or introduce significant latency. We propose Style Arithmetic (SA), a novel parameter-space approach that first extracts style-specific representations by analyzing parameter differences between models trained on contrasting styles, then incorporates these representations into a base model with precise control over style intensity. Our experiments show that SA achieves three key capabilities: controllability for precise adjustment of styles, transferability for effective style transfer across tasks, and composability for simultaneous control of multiple style dimensions. Compared to alternative methods, SA offers superior effectiveness while achieving optimal computational efficiency. Our approach opens new possibilities for flexible and efficient style control in language models.

Dynamic Order Template Prediction for Generative Aspect-Based Sentiment Analysis

Yonghyun Jun, Hwanhee Lee

Aspect-based sentiment analysis (ABSA) assesses sentiments towards specific aspects within texts, resulting in detailed sentiment tuples. Previous ABSA models often used static templates to predict all the elements in the tuples, and these models often failed to accurately capture dependencies between elements. Multi-view prompting method improves the performance of ABSA by predicting tuples with various templates and then assembling the results. However, this method suffers from inefficiencies and out-of-distribution errors. In this paper, we propose a Dynamic Order Template (DOT) method for ABSA, which dynamically creates an order template that contains only the necessary views for each instance. Ensuring the diverse and relevant view generation, our proposed method improves F1 scores on ASQP and ACOS datasets while significantly reducing inference time.

PVP: An Image Dataset for Personalized Visual Persuasion with Persuasiveness Ratings, Persuasion Strategies, and Viewer Characteristic

Junseo Kim, Jongwook Han, Dongmin Choi, Jongwook Yoon, Eun-Ju Lee, Yohan Jo

Visual persuasion, which uses visual elements to influence cognition and behaviors, is crucial in fields such as advertising and political communication. With recent advancements in artificial intelligence, there is growing potential to develop persuasive systems that automatically generate persuasive images tailored to individuals. However, a significant bottleneck in this area is the lack of comprehensive datasets that connect the persuasiveness of images with the personal information about those who evaluated the images. To address this gap and facilitate technological advancements in personalized visual persuasion, we release the Personalized Visual Persuasion (PVP) dataset, comprising 28,454 persuasive images across 596 messages and 9 persuasion strategies. Importantly, the PVP dataset provides persuasiveness scores of images evaluated by 2,521 human annotators, along with their demographic and psychological characteristics (personality traits and values). We demonstrate the utility of our dataset by developing a persuasive image generator and an automated evaluator, and establish benchmark baselines. Our experiments reveal that incorporating psychological characteristics enhances the generation and evaluation of persuasive images, providing valuable insights for personalized visual persuasion.

CU-MAM: Coherence-Driven Unified Macro-Structures for Argument Mining

Debela Gemechu, Chris Reed

Argument Mining (AM) involves the automatic identification of argument structure in natural language. Traditional AM methods rely on micro-structural features derived from the internal properties of individual Argumentative Discourse Units (ADUs). However, argument structure is shaped by a macro-structure capturing the functional interdependence among ADUs. This macro-structure consists of segments, where each segment contains ADUs that fulfill specific roles to maintain coherence within the segment (**local coherence**) and across segments (**global coherence**). This paper presents an approach that models macro-structure, capturing both local and global coherence to identify argument structures. Experiments on heterogeneous datasets demonstrate superior performance in both in-dataset and cross-dataset evaluations. The cross-dataset evaluation shows that macro-structure enhances transferability to unseen datasets.

PerSphere: A Comprehensive Framework for Multi-Faceted Perspective Retrieval and Summarization

Yun Luo, Yingjie Li, Xiangkun Hu, Qinglin Qi, Fang Guo, Qipeng Guo, Zheng Zhang, Yue Zhang

As online platforms and recommendation algorithms evolve, people are increasingly trapped in echo chambers, leading to biased understandings of various issues. To combat this issue, we have introduced PerSphere, a benchmark designed to facilitate multi-faceted perspective retrieval and summarization, thus breaking free from these information silos. For each query within PerSphere, there are two opposing claims, each supported by distinct, non-overlapping perspectives drawn from one or more documents. Our goal is to accurately summarize these documents, aligning the summaries with the respective claims and their underlying perspectives. This task is structured as a two-step end-to-end pipeline that includes comprehensive document retrieval and multi-faceted summarization. Furthermore, we propose a set of metrics to evaluate the comprehensiveness of the retrieval and summarization content. Experimental results on various counterparts for the pipeline show that recent models struggle with such a complex task. Analysis shows that the main challenge lies in long context and perspective extraction, and we propose a simple but effective multi-agent summarization system, offering a promising solution to enhance performance on PerSphere.

Does the Emotional Understanding of LVLMs Vary Under High-Stress Environments and Across Different Demographic Attributes?

Jaewook Lee, Yeajin Jang, Oh-Woog KWON, Harksoo Kim

According to psychological and neuroscientific research, a high-stress environment can restrict attentional resources and intensify negative affect, thereby impairing the ability to understand emotions. Furthermore, demographic attributes such as race, gender, and age group have been repeatedly reported to cause significant differences in emotional expression and recognition. This study is the first to systematically verify whether these psychological findings observed in humans also apply to the latest Large Vision Language Models (LVLMs). We constructed low-stress versus high-stress environments and generated an image dataset (a total of 540 images) that combines race, gender, and age group. Based on this, we applied the Pretend prompt technique to induce LVLMs to interpret others' emotions from the standpoint of the assigned environment and persona. An analysis of the models' emotional understanding ability, using EQ-Bench-based metrics, revealed that (1) under high-stress environments, the accuracy of emotion understanding significantly declined in most LVLMs, and (2) performance disparities were confirmed across race, gender, and age group. These findings suggest that the effects of high-stress and demographic attributes identified in human research may also be reflected in LVLMs.

CoE: A Clue of Emotion Framework for Emotion Recognition in Conversations

Zhiyu Shen, Yunhe Pang, Yanghui Rao, Jianxing Yu

Emotion Recognition in Conversations (ERC) is crucial for machines to understand dynamic human emotions. While Large Language Models (LLMs) show promise, their performance is often limited by challenges in interpreting complex conversational streams. We introduce a Clue of Emotion (CoE) framework, which progressively integrates key conversational clues to enhance the ERC task. Building on CoE, we implement a multi-stage auxiliary learning strategy that incorporates role-playing, speaker identification, and emotion reasoning tasks, each targeting different aspects of conversational emotion understanding and enhancing the model's ability to interpret emotional contexts. Our experiments on EmoryNLP, MELD, and IEMOCAP demonstrate that CoE consistently outperforms state-of-the-art methods, achieving a 2.92\% improvement on EmoryNLP. These results underscore the effectiveness of clues and multi-stage auxiliary learning for ERC, offering valuable insights for future research.

Colloquial Singaporean English Style Transfer with Fine-Grained Explainable Control

Jinggui Liang, Dung Vo, Yap Hong Xian, Hai Leong Chieu, Kian Ming A. Chai, Jing Jiang, Lizi Liao

Colloquial Singaporean English (Singlish) is an informal English marked by a unique blend of languages reflecting Singapore's multicultural identity. Style transfer between Singlish and Standard (formal) English is vital for various applications, yet existing methods often lack explainability and fine-grained control. To fill this gap, we contribute in two key ways. First, we construct a large, high-quality dataset of formal and informal sentences, annotated across six linguistic aspects—Syntax, Lexical Borrowing, Pragmatics, Prosody/Phonology, Emoticons/Punctuation, and Code-Switching—with detailed explanations. Starting with manually annotated cases, we scaled the dataset to 140K with ensured quality. Second, inspired by the "Society of Mind" theory, we propose a novel multi-agent framework where large language models (LLMs) act as expert agents for each linguistic aspect. These agents collaborate by iteratively generating, critiquing, and refining responses to achieve controlled, explainable style transfer. Both automatic metrics and human evaluations confirm that our method enables precise, interpretable transformations, advancing explainability in NLP for Singlish.

InspireDebate: Multi-Dimensional Subjective-Objective Evaluation-Guided Reasoning and Optimization for Debating

Fuyu Wang, Jiangtong Li, Kun Zhu, Changjun Jiang

With the rapid advancements in large language models (LLMs), debating tasks, such as argument quality assessment and debate process simulation, have made significant progress. However, existing LLM-based debating systems focus on responding to specific arguments while neglecting objective assessments such as authenticity and logical validity. Furthermore, these systems lack a structured approach to optimize across various dimensions—including evaluation metrics, chain-of-thought (CoT) reasoning, and multi-turn debate refinement—thereby limiting their effectiveness. To address these interconnected challenges, we propose a dual-component framework: (1) InspireScore, a novel evaluation system that establishes a multi-dimensional assessment architecture incorporating four subjective criteria (emotional appeal, argument clarity, argument arrangement, and topic relevance) alongside two objective metrics (fact authenticity and logical validity); and (2) InspireDebate, an optimized debating framework employing a phased optimization approach through CoT reasoning enhancement, multi-dimensional Direct Preference Optimization (DPO), and real-time knowledge grounding via web-based Retrieval Augmented Generation (Web-RAG). Empirical evaluations demonstrate that InspireScore achieves 44% higher correlation with expert judgments compared to existing methods, while InspireDebate shows significant improvements, outperforming baseline models by 57%. Source code is available at https://github.com/fywang12/InspireDebate.

Session 2: Speech recognition, text-to-speech and spoken language understanding (Monday 28 July, 11:00-12:30, Hall 4/5)

Autoregressive Speech Synthesis without Vector Quantization

Lingwei Meng, Long Zhou, Shujie LIU, Sanyuan Chen, Bing Han, Shujie HU, Yanqing Liu, Jinyu Li, sheng zhao, Xixin Wu, Helen M. Meng, Furu Wei

We present MELLE, a novel continuous-valued token based language modeling approach for text-to-speech synthesis (TTS). MELLE autoregressively generates continuous mel-spectrogram frames directly from text condition, bypassing the need for vector quantization, which is typically designed for audio compression and sacrifices fidelity compared to continuous representations. Specifically, (i) instead of cross-entropy loss, we apply regression loss with a proposed spectrogram flux loss function to model the probability distribution of the continuous-valued tokens; (ii) we have incorporated variational inference into MELLE to facilitate sampling mechanisms, thereby enhancing the output diversity and model robustness. Experiments demonstrate that, compared to the two-stage codec language model VALL-E and its variants, the single-stage MELLE mitigates robustness issues by avoiding the inherent flaws of sampling vector-quantized codes, achieves superior performance across multiple metrics, and, most importantly, offers a more streamlined paradigm. The demos of our work are provided at https://aka.ms/melle.

F5-TTS: A Fairytaler that Fakes Fluent and Faithful Speech with Flow Matching

Yushen CHEN, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao, Kai Yu, Xie Chen

This paper introduces F5-TTS, a fully non-autoregressive text-to-speech system based on flow matching with Diffusion Transformer (DiT). Without requiring complex designs such as duration model, text encoder, and phoneme alignment, the text input is simply padded with filler tokens to the same length as input speech, and then the denoising is performed for speech generation, which was originally proved feasible by E2 TTS. However, the original design of E2 TTS makes it hard to follow due to its slow convergence and low robustness. To address these issues, we first model the input with ConvNeXt to refine the text representation, making it easy to align with the speech. We further propose an inference-time Sway Sampling strategy, which significantly improves our model's performance and efficiency. This sampling strategy for flow step can be easily applied to existing flow matching based models without retraining. Our design allows faster training and achieves an inference RTF of 0.15, which is greatly improved compared to state-of-the-art diffusion-based TTS models. Trained on a public 100K hours multilingual dataset, our F5-TTS exhibits highly natural and expressive zero-shot ability, seamless code-switching capability, and speed control efficiency. We have released all codes and checkpoints to promote community development, at https://SWivid.github.io/F5-TTS/.

SIFT-50M: A Large-Scale Multilingual Dataset for Speech Instruction Fine-Tuning

Prabhat Pandey, Rupak Vignesh Swaminathan, K V Vijay Girish, Arunasish Sen, Jian. Xie, Grant Strimel, Andreas Schwarz

We introduce SIFT (Speech Instruction Fine-Tuning), a 50M-example dataset designed for instruction fine-tuning and pre-training of speech-text large language models (LLMs). SIFT-50M is built from publicly available speech corpora, which collectively contain 14K hours of speech, and leverages LLMs along with off-the-shelf expert models. The dataset spans five languages, encompassing a diverse range of speech understanding as well as controllable speech generation instructions. Using SIFT-50M, we train SIFT-LLM, which outperforms existing speech-text LLMs on instruction-following benchmarks while achieving competitive performance on foundational speech tasks. To support further research, we also introduce EvalSIFT, a benchmark dataset specifically designed to evaluate the instruction-following capabilities of speech-text LLMs.

Recent Advances in Speech Language Models: A Survey

Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Steven Y. Guo, Irwin King

Text-based Large Language Models (LLMs) have recently gained significant attention, primarily for their capabilities in text-based interactions. However, natural human interaction often relies on speech, highlighting the need for voice-based models. In this context, Speech Language Models (SpeechLMs)---foundation models designed to understand and generate speech---emerge as a promising solution for end-to-end speech interaction. This survey offers a comprehensive overview of recent approaches to building SpeechLMs, outlining their core architectural components, training methodologies, evaluation strategies, and the challenges and potential directions for future research in this rapidly advancing field. The GitHub repository is available at https://github.com/dreamtheater123/Awesome-SpeechLM-Survey

SimulS2S-LLM: Unlocking Simultaneous Inference of Speech LLMs for Speech-to-Speech Translation

Keqi Deng, Wenxi Chen, Xie Chen, Phil Woodland

Simultaneous speech translation (SST) outputs translations in parallel with streaming speech input, balancing translation quality and latency. While large language models (LLMs) have been extended to handle the speech modality, streaming remains challenging as speech is pre-pended as a prompt for the entire generation process. To unlock LLM streaming capability, this paper proposes SimulS2S-LLM, which trains speech LLMs offline and employs a test-time policy to guide simultaneous inference. SimulS2S-LLM alleviates the mismatch between training and inference by extracting boundary-aware speech prompts that allows it to be better matched with text input data. SimulS2S-LLM achieves simultaneous speech-to-speech translation (Simul-S2ST) by predicting discrete output speech tokens and then synthesising output speech using a pre-trained vocoder. An incremental beam search is designed to expand the search space of speech token prediction without increasing latency. Experiments on the CVSS speech data show that SimulS2S-LLM offers a better translation quality-latency trade-off than existing methods that use the same training data, such as improving ASR-BLEU scores by 3 points at similar latency.

VoxEval: Benchmarking the Knowledge Understanding Capabilities of End-to-End Spoken Language Models

Wenqian Cui, Xiaoqi Jiao, Ziqiao Meng, Irwin King

With the rising need for speech-based interaction models, end-to-end Spoken Language Models (SLMs) have emerged as a promising solution. While these models require comprehensive world knowledge for meaningful and reliable human interactions, existing question-answering (QA) benchmarks fall short in evaluating SLMs' knowledge understanding due to their inability to support end-to-end speech evaluation and account for varied input audio conditions. To address these limitations, we present VoxEval, a novel SpeechQA benchmark that assesses SLMs' knowledge understanding through pure speech interactions. Our benchmark uniquely maintains speech format for both inputs and outputs, evaluates model robustness across diverse input audio conditions, and pioneers the assessment of complex tasks like mathematical reasoning in spoken format. Through systematic evaluation, we demonstrate that current SLMs exhibit poor performance on VoxEval, show sensitivity to varying audio conditions, and possess limited reasoning capabilities, highlighting critical areas for future development. VoxEval dataset is available at: https://github.com/dreamtheater123/VoxEval

InSerter: Speech Instruction Following with Unsupervised Interleaved Pre-training

Dingdong WANG, Jin Xu, Ruihang Chu, Zhifang Guo, Xiong Wang, Jincenzi Wu, Dongchao Yang, Shengpeng Ji, Junyang Lin

Recent advancements in speech large language models (SpeechLLMs) have attracted considerable attention. Nonetheless, current methods exhibit suboptimal performance in adhering to speech instructions. Notably, the intelligence of models significantly diminishes when processing speech-form input as compared to direct text-form input. Prior work has attempted to mitigate this semantic inconsistency between speech and text representations through techniques such as representation and behavior alignment, which involve the meticulous design of data pairs during the post-training phase. In this paper, we introduce a simple and scalable training method called InSerter, which stands for Interleaved Speech-Text Representation Pre-training. InSerter is designed to pre-train large-scale unsupervised speech-text sequences, where the speech is synthesized from randomly selected segments of an extensive text corpus using text-to-speech conversion. Consequently, the model acquires the ability to generate textual continuations corresponding to the provided speech segments, obviating the need for intensive data design endeavors. To systematically evaluate speech instruction-following capabilities, we introduce SpeechInstructBench, the first comprehensive benchmark specifically designed for speech-oriented instruction-following tasks. Our proposed model InSerter achieves SOTA performance in SpeechInstructBench and demonstrates superior or competitive results across diverse speech processing tasks.

DNCASR: End-to-End Training for Speaker-Attributed ASR

Xianrui Zheng, Chao Zhang, Phil Woodland

This paper introduces DNCASR, a novel end-to-end trainable system designed for joint neural speaker clustering and automatic speech recognition (ASR), enabling speaker-attributed transcription of long multi-party meetings. DNCASR uses two separate encoders to independently encode global speaker characteristics and local waveform information, along with two linked decoders to generate speaker-attributed transcriptions. The use of linked decoders allows the entire system to be jointly trained under a unified loss function. By employing a serialised training approach, DNCASR effectively addresses overlapping speech in real-world meetings, where the link improves the prediction of speaker indices in overlapping segments. Experiments on the AMI-MDM meeting corpus demonstrate that the jointly trained DNCASR outperforms a parallel system that does not have links between the speaker and ASR decoders. Using cpWER to measure the speaker-attributed word error rate, DNCASR achieves a 9.0% relative reduction on the AMI-MDM Eval set.

LLaMA-Omni 2: LLM-based Real-time Spoken Chatbot with Autoregressive Streaming Speech Synthesis

Qingkai Fang, Yan Zhou, Shoutao Guo, Shaolei Zhang, Yang Feng

Real-time, intelligent, and natural speech interaction is an essential part of the next-generation human-computer interaction. Recent advancements have showcased the potential of building intelligent spoken chatbots based on large language models (LLMs). In this paper, we introduce LLaMA-Omni 2, a series of speech language models (SpeechLMs) ranging from 0.5B to 14B parameters, capable of achieving high-quality real-time speech interaction. LLaMA-Omni 2 is built upon the Qwen2.5 series models, integrating a speech encoder and an autoregressive streaming speech decoder. Despite being trained on only 200K multi-turn speech dialogue samples, LLaMA-Omni 2 demonstrates strong performance on several spoken question answering and speech instruction following benchmarks, surpassing previous state-of-the-art SpeechLMs like GLM-4-Voice, which was trained on millions of hours of speech data.

UniCodec: Unified Audio Codec with Single Domain-Adaptive Codebook

Yidi Jiang, Qian Chen, Shengpeng Ji, Yu Xi, Wen Wang, Chong Zhang, Xianghu Yue, ShiLiang Zhang, Haizhou Li

The emergence of audio language models is empowered by neural audio codecs, which establish critical mappings between continuous waveforms and discrete tokens compatible with language model paradigms. The evolutionary trends from multi-layer residual vector quantizer to single-layer quantizer are beneficial for language-autoregressive decoding. However, the capability to handle multi-domain audio signals through a single codebook remains constrained by inter-domain distribution discrepancies. In this work, we introduce UniCodec, a unified audio codec with a single codebook to support multi-domain audio data, including speech, music, and sound. To achieve this, we propose a partitioned domain-adaptive codebook method based on domain Mixture-of-Experts strategy to capture the distinct characteristics of each audio domain. Furthermore, to enrich the semantic density of the codec without auxiliary modules, we propose a self-supervised mask prediction modeling approach. Comprehensive objective and subjective evaluations demonstrate that UniCodec achieves excellent audio reconstruction performance across the three audio domains, outperforming existing unified neural codecs with a single codebook, and even surpasses state-of-the-art domain-specific codecs on both acoustic and semantic representation capabilities.

ZIPA: A family of efficient models for multilingual phone recognition

Jian Zhu, Farhan Samir, Eleanor Chodroff, David R. Mortensen

We present ZIPA, a family of efficient speech models that advances the state-of-the-art performance of crosslinguistic phone recognition. We first curated IPA PACK++, a large-scale multilingual speech corpus with 17,000+ hours of normalized phone transcriptions and a novel evaluation set capturing unseen languages and sociophonetic variation. ZIPA, including transducer (ZIPA-T) and CTC-based (ZIPA-CR) variants, leverages the efficient Zipformer backbones and outperforms existing phone recognition systems with much fewer parameters. Further scaling via noisy student training on 11,000+ hours of pseudo-labeled multilingual data yields further improvement. While ZIPA achieves strong performance on benchmarks, error analysis reveals persistent limitations in modeling sociophonetic diversity, underscoring challenges for future research.

Align-SLM: Textless Spoken Language Models with Reinforcement Learning from AI Feedback

Guan-Ting Lin, Prashanth Gurunath Shivakumar, Aditya Gourav, Yile Gu, Ankur Gandhe, Hung-yi Lee, Ivan Bulyko

While textless Spoken Language Models (SLMs) have shown potential in end-to-end speech-to-speech modeling, they still lag behind text-based Large Language Models (LLMs) in terms of semantic coherence and relevance. This work introduces the \textbf{Align-SLM} framework, which leverages preference optimization inspired by Reinforcement Learning with Human Feedback (RLHF) to enhance the semantic understanding of SLMs. Our approach generates multiple speech continuations from a given prompt and uses LLM-based semantic metrics to create preference data for Direct Preference Optimization (DPO). We evaluate the framework using ZeroSpeech 2021 benchmarks for lexical and syntactic modeling, the spoken version of the StoryCloze dataset for semantic coherence, and other speech generation metrics, including the GPT4-o score and human evaluation. Experimental results show that our method achieves the state-of-the-art performance of SLMs for most benchmarks, highlighting the importance of preference optimization to improve the semantics of SLMs.

OZSpeech: One-step Zero-shot Speech Synthesis with Learned-Prior-Conditioned Flow Matching

Nghia Huynh Nguyen Hieu, Ngoc Son Nguyen, Huynh Nguyen Dang, Thieu Vo, Truong-Son Hy, Van Nguyen

Text-to-speech (TTS) systems have seen significant advancements in recent years, driven by improvements in deep learning and neural network architectures. Viewing the output speech as a data distribution, previous approaches often employ traditional speech representations, such as waveforms or spectrograms, within the Flow Matching framework. However, these methods have limitations, including overlooking various speech attributes and incurring high computational costs due to additional constraints introduced during training. To address these challenges, we introduce OZSpeech, the first TTS method to explore optimal transport conditional flow matching with one-step sampling and a learned prior as the condition, effectively disregarding preceding states and reducing the number of sampling steps. Our approach operates on disentangled, factorized components of speech in token format, enabling accurate modeling of each speech attribute, which enhances the TTS system's ability to precisely clone the prompt speech. Experimental results show that our method achieves promising performance over existing methods in content accuracy, naturalness, prosody generation, and speaker style preservation. Audio samples are available at our demo page https://ozspeech.github.io/OZSpeech_Web/.

ECLM: Entity Level Language Model for Spoken Language Understanding with Chain of Intent

Shangjian Yin, Peijie Huang, JiaTian Chen, Haojing Huang, Yuhong Xu

Large Language Models (LLMs) have demonstrated impressive capabilities in language generation and general task performance. However, their application to spoken language understanding (SLU) remains challenging, particularly for token-level tasks, where the autoregressive nature of LLMs often leads to misalignment issues. They also struggle to capture nuanced interrelations in semantic-level tasks through direct fine-tuning alone. To address these challenges, we propose the Entity-level Language Model (ECLM) framework, which reformulates slot-filling as an entity recognition task and introduces a novel concept, Chain of Intent, to enable step-by-step multi-intent recognition. Experimental results show that ECLM significantly outperforms strong baselines such as Uni-MIS, achieving gains of 3.7% on MixATIS and 3.1% on MixSNIPS. Compared to standard supervised fine-tuning of LLMs, ECLM further achieves improvements of 8.5% and 21.2% on these datasets, respectively. Our code is available at https://github.com/SJY8460/ECLM.

Measuring the Effect of Transcription Noise on Downstream Language Understanding Tasks

Ori Shapira, Shlomo Chazan, Amir David Nissan Cohen

With the increasing prevalence of recorded human speech, spoken language understanding (SLU) is essential for its efficient processing. In order to process the speech, it is commonly transcribed using automatic speech recognition technology. This speech-to-text transition introduces errors into the transcripts, which subsequently propagate to downstream NLP tasks, such as dialogue summarization. While it is known that transcript noise affects downstream tasks, a general-purpose and systematic approach to analyzing its effects across different noise severities and types has not been addressed. We propose a configurable framework for assessing task models in diverse noisy settings, and for examining the impact of transcript-cleaning techniques. The framework facilitates the investigation of task model behavior, which can in turn support the development of effective SLU solutions. We exemplify the utility of our framework on three SLU tasks and four task models, offering insights regarding the effect of transcript noise on tasks in general and models in particular. For instance, we find that task models can tolerate a certain level of noise, and are affected differently by the types of errors in the transcript.

SpeechIQ: Speech-Agentic Intelligence Quotient Across Cognitive Levels in Voice Understanding by Large Language Models

Zhen Wan, Chao-Han Huck Yang, Yahan Yu, Jinchuan Tian, Sheng Li, Ke Hu, Zhehuai Chen, Shinji Watanabe, Fei Cheng, Chenhui Chu, Yu-Chiang Frank Wang, Sadao Kurohashi

We introduce Speech-based Intelligence Quotient (SIQ) as a new form of human cognition-inspired evaluation pipeline for voice understanding large language models (LLM_{Voice}), designed to assess their voice understanding ability. Moving beyond popular voice understanding metrics such as word error rate (WER), SIQ examines LLM_{Voice} across three cognitive levels motivated by Bloom's Taxonomy: (1) Remembering (i.e., WER for verbatim accuracy); (2) Understanding (i.e., similarity of LLM's interpretations); and (3) Application (i.e., QA accuracy for simulating downstream tasks). We demonstrate that SIQ not only quantifies voice understanding abilities but also provides unified comparisons between cascaded methods (e.g., ASR-LLM) and end-to-end models, identifies annotation errors in existing benchmarks, and detects hallucinations in LLM_{Voice}. Our framework represents a first-of-its-kind intelligence examination that bridges cognitive principles with voice-oriented benchmarks, while exposing overlooked challenges in multi-modal training. Our code and data will be open source to encourage future studies.

Session 2: Speech recognition, text-to-speech and spoken language understanding (Monday 28 July, 11:00-12:30, Hall 5X)

MERaLiON-AudioLLM: Advancing Speech and Language Understanding for Singapore

Yingxu He, Zhuohan Liu, Geyu Lin, Shuo Sun, Bin Wang, Wenyu Zhang, Xunlong Zou, Nancy F. Chen, AiTi Aw

We introduce MERaLiON-AudioLLM, the first general-purpose audio-based large language model designed for multitask learning, with a particular focus on Singlish understanding. Trained on 62 million multimodal instruction samples comprising a total of 260k hours of audio, it exhibits strong generalization across a diverse set of tasks, including—but not limited to—automatic speech recognition, spoken question answering, speech translation, and paralinguistic analysis. Our results show significant improvements in local speech recognition and task-specific understanding, making MERaLiON-AudioLLM a leading solution for region-specific AI applications. An interactive demo has been developed to enable user-friendly interactions, supported by a backend with customized caching and load-balancing mechanisms. We benchmark the model across a broad range of multilingual and multitask scenarios, where it demonstrates competitive performance compared to other open-source models. The demo page, model weights and videos are publically accessible.

RT-VC: Real-Time Zero-Shot Voice Conversion with Speech Articulatory Coding

Yisi Liu, Chenyang Wang, Hanjo Kim, Raniya Khan, Gopala Anumanchipalli

Voice conversion has emerged as a pivotal technology in numerous applications ranging from assistive communication to entertainment. In this paper, we present RT-VC, a zero-shot real-time voice conversion system that delivers ultra-low latency and high-quality performance. Our approach leverages an articulatory feature space to naturally disentangle content and speaker characteristics, facilitating more robust and interpretable voice transformations. Additionally, the integration of differentiable digital signal processing (DDSP) enables efficient vocoding directly from articulatory features, significantly reducing conversion latency. Experimental evaluations demonstrate that, while maintaining synthesis quality comparable to the current state-of-the-art (SOTA) method, RT-VC achieves a CPU latency of 61.4 ms, representing a 13.3% reduction in latency. Project page: https://berkeley-speech-group.github.io/RT-VC/

Session 2: Summarization (Monday 28 July, 11:00-12:30, Hall 4/5)

What Is That Talk About? A Video-to-Text Summarization Dataset for Scientific Presentations

Dongqi Liu, Chenxi Whitehouse, Xi Yu, Louis Mahon, Rohit Saxena, Zheng Zhao, Yifu QIU, Mirella Lapata, Vera Demberg

Transforming recorded videos into concise and accurate textual summaries is a growing challenge in multimodal learning. This paper introduces VISTA, a dataset specifically designed for video-to-text summarization in scientific domains. VISTA contains 18,599 recorded AI conference presentations paired with their corresponding paper abstracts. We benchmark the performance of state-of-the-art large models and apply a plan-based framework to better capture the structured nature of abstracts. Both human and automated evaluations confirm that explicit planning enhances summary quality and factual consistency. However, a considerable gap remains between models and human performance, highlighting the challenges of our dataset. This study aims to pave the way for future research on scientific video-to-text summarization.

NexusSum: Hierarchical LLM Agents for Long-Form Narrative Summarization

Hyuntak Kim, Byung-Hak Kim

Summarizing long-form narratives—such as books, movies, and TV scripts—requires capturing intricate plotlines, character interactions, and thematic coherence, a task that remains challenging for existing LLMs. We introduce NexusSum, a multi-agent LLM framework for narrative summarization that processes long-form text through a structured, sequential pipeline—without requiring fine-tuning. Our approach introduces two key innovations: **(1) Dialogue-to-Description Transformation**: A narrative-specific preprocessing method that standardizes character dialogue and descriptive text into a unified format, improving coherence. **(2) Hierarchical Multi-LLM Summarization**: A structured summarization pipeline that optimizes chunk processing and controls output length for accurate, high-quality summaries. Our method establishes a new state-of-the-art in narrative summarization, achieving up to **a 30.0% improvement in BERTScore (F1)** across books, movies, and TV scripts. These results demonstrate the effectiveness of multi-agent LLMs in handling long-form content, offering a scalable approach for structured summarization in diverse storytelling domains.

Towards Multi-dimensional Evaluation of LLM Summarization across Domains and Languages

Hyangsuk Min, Yuho Lee, Minjeong Ban, Jiaqi Deng, Nicole Hee-Yeon Kim, Taewon Yun, Hang Su, Jason Cai, Hwanjun Song

Evaluation frameworks for text summarization have evolved in terms of both domain coverage and metrics. However, existing benchmarks still lack domain-specific assessment criteria, remain predominantly English-centric, and face challenges with human annotation due to the complexity of reasoning. To address these, we introduce MSumBench, which provides a multi-dimensional, multi-domain evaluation of summarization in English and Chinese. It also incorporates specialized assessment criteria for each domain and leverages a multi-agent debate system to enhance annotation quality. By evaluating eight modern summarization models, we discover distinct performance patterns across domains and languages. We further examine large language models as summary evaluators, analyzing the correlation between their evaluation and summarization capabilities, and uncovering systematic bias in their assessment of self-generated summaries. Our benchmark dataset is publicly available at https://github.com/DISL-Lab/MSumBench.

Enhancing Event-centric News Cluster Summarization via Data Sharpening and Localization Insights

Longyin Zhang, Bowei Zou, AiTi Aw

This paper tackles the challenges of clustering news articles by main events (MEs) and summarizing these clusters, focusing on diverse languages and localized contexts. Our approach consists of four key contributions. First, we investigate the role of dynamic clustering and the integration of various ME references, including event attributions extracted by language models (LMs), in enhancing event-centric clustering. Second, we propose a data-sharpening framework that optimizes the balance between information volume and entropy in input texts, thereby optimizing generated summaries on multiple indicators. Third, we fine-tune LMs with local news articles for cross-lingual temporal question-answering and text summarization, achieving notable improvements in capturing localized contexts. Lastly, we present the first cross-lingual dataset and comprehensive evaluation metrics tailored for the event-centric news cluster summarization pipeline. Our findings enhance the understanding of news summarization across N-gram, event-level coverage, and faithfulness, providing new insights into leveraging LMs for large-scale cross-lingual and localized news analysis.

QQSUM: A Novel Task and Model of Quantitative Query-Focused Summarization for Review-based Product Question Answering

An Quang Tang, Xiuzhen Zhang, Minh Ngoc Dinh, Zhuang Li

Review-based Product Question Answering (PQA) allows e-commerce platforms to automatically address customer queries by leveraging insights from user reviews. However, existing PQA systems generate answers with only a single perspective, failing to capture the diversity of customer opinions. In this paper we introduce a novel task Quantitative Query-Focused Summarization (QQSUM), which aims to summarize diverse customer opinions into representative Key Points (KPs) and quantify their prevalence to effectively answer user queries. While Retrieval-Augmented Generation (RAG) shows promise for PQA, its generated answers still fall short of capturing the full diversity of viewpoints. To tackle this challenge, our model QQSUM-RAG, which extends RAG, employs few-shot learning to jointly train a KP-oriented retriever and a KP summary generator, enabling KP-based summaries that capture diverse and representative opinions. Experimental results demonstrate that QQSUM-RAG achieves superior performance compared to state-of-the-art RAG baselines in both textual quality and quantification accuracy of opinions. Our source code is available at: https://github.com/antangrocket1312/QQSUMM

Temporal reasoning for timeline summarisation in social media

Jiayu Song, Mahmud Elahi Akhter, Dana Atzil-Slonim, Maria Liakata

This paper explores whether enhancing temporal reasoning capabilities in Large Language Models (LLMs) can improve the quality of timeline summarisation, the task of summarising long texts containing sequences of events, such as social media threads. We first introduce NarrativeReason, a novel dataset focused on temporal relationships among sequential events within narratives, distinguishing it from existing temporal reasoning datasets that primarily address pair-wise event relationships. Our approach then combines temporal reasoning with timeline summarisation through a knowledge distillation framework, where we first fine-tune a teacher model on temporal reasoning tasks and then distill this knowledge into a student model while simultaneously training it for the task of timeline summarisation. Experimental results demonstrate that our model achieves superior performance on out-of-domain mental health-related timeline summarisation tasks, which involve long social media threads with repetitions of events and a mix of emotions, highlighting the importance and generalisability of leveraging temporal reasoning to improve timeline summarisation.

Principled Content Selection to Generate Diverse and Personalized Multi-Document Summaries

Vishakh Padmakumar, Zichao Wang, David Arbour, Jennifer Healey

While large language models (LLMs) are increasingly capable of handling longer contexts, recent work has demonstrated that they exhibit the _"lost in the middle"_ phenomenon (Liu et al., 2024) of unevenly attending to different parts of the provided context. This hinders their ability to cover diverse source material in multi-document summarization, as noted in the DiverseSumm benchmark (Huang et al., 2024). In this work, we contend that principled content selection is a simple way to increase source coverage on this task. As opposed to prompting an LLM to perform the summarization in a single step, we explicitly divide the task into three steps---(1) reducing document collections to atomic key points, (2) using determinantal point processes (DPP) to perform select key points that prioritize diverse content, and (3) rewriting to the final summary. By combining prompting steps, for extraction and rewriting, with principled techniques, for content selection, we consistently improve source coverage on the DiverseSumm benchmark across various LLMs. Finally, we also show that by incorporating relevance to a provided user intent into the DPP kernel, we can generate _personalized_ summaries that cover _relevant_ source information while retaining coverage.

Session 2: TACL Journal (Monday 28 July, 11:00-12:30, Hall 4/5)

Retrieval-style In-Context Learning for Few-shot Hierarchical Text Classification

Huiyao Chen, Zulong Chen, Yu Zhao, Mengjia Wang, Liangyue Li, Meishan Zhang, Min Zhang

Hierarchical text classification (HTC) is an important task with broad applications, while few-shot HTC has gained increasing interest recently. While in-context learning (ICL) with large language models (LLMs) has achieved significant success in few-shot learning, it is not as effective for HTC because of the expansive hierarchical label sets and extremely-ambiguous labels. In this work, we introduce the first ICL-based framework with LLM for few-shot HTC. We exploit a retrieval database to identify relevant demonstrations, and an iterative policy to manage multi-layer hierarchical labels. Particularly, we equip the retrieval database with HTC label-aware representations for the input texts, which is achieved by continual training on a pretrained language model with masked language modeling (MLM), layer-wise classification (CLS, specifically for HTC), and a novel divergent contrastive learning (DCL, mainly for adjacent semantically-similar labels) objective. Experimental results on three benchmark datasets demonstrate superior performance of our method, and we can achieve state-of-the-art results in few-shot HTC.

Hypernetworks for Personalizing ASR to Atypical Speech

Dianna Yee, Max Müller-Eberstein, Karren Yang, Gautam Varma Mantena, Colin Lea

Parameter-efficient fine-tuning (PEFT) for personalizing automatic speech recognition (ASR) has recently shown promise for adapting general population models to atypical speech. However, these approaches assume a priori knowledge of the atypical speech disorder being adapted for—the diagnosis of which requires expert knowledge that is not always available. Even given this knowledge, data scarcity and high inter/intra-speaker variability further limit the effectiveness of traditional fine-tuning. To circumvent these challenges, we first identify the minimal set of model parameters required for ASR adaptation. Our analysis of each individual parameter's effect on adaptation performance allows us to reduce Word Error Rate (WER) by half while adapting 0.03% of all weights. Alleviating the need for cohort-specific models, we next propose the novel use of a meta-learned hypernetwork to generate highly individualized, utterance-level adaptations on-the-fly for a diverse set of atypical speech characteristics. Evaluating adaptation at the global, cohort and individual-level, we show that hypernetworks generalize better to out-of-distribution speakers, while maintaining an overall relative WER reduction of 75.2% using 0.1% of the full parameter budget.

CAMELL: Confidence-based Acquisition Model for Efficient Self-supervised Active Learning with Label Validation

Carel van Niekerk, Christian Geishauser, Michael Heck, Shutong Feng, Hsien-Chin Lin, Nurul Lubis, Benjamin Ruppik, Renato Vukovic, Milica Gašić

Supervised neural approaches are hindered by their dependence on large, meticulously annotated datasets, a requirement that is particularly cumbersome for sequential tasks. The quality of annotations tends to deteriorate with the transition from expert-based to crowd-sourced labelling. To address these challenges, we present CAMELL (Confidence-based Acquisition Model for Efficient self-supervised active Learning with Label validation), a pool-based active learning framework tailored for sequential multi-output problems. CAMELL possesses three core features: (1) it requires expert annotators to label only a fraction of a chosen sequence, (2) it facilitates self-supervision for the remainder of the sequence, and (3) it employs a label validation mechanism to prevent erroneous labels from contaminating the dataset and harming model performance. We evaluate CAMELL on two sequential tasks, with a special emphasis on dialogue belief tracking, a task plagued by the constraints of limited and noisy datasets. Our experiments demonstrate that CAMELL outperforms the baselines in terms of efficiency. Furthermore, the data corrections suggested by our method contribute to an overall improvement in the quality of the resulting datasets.

Phonetic Reconstruction of the Consonant System of Middle Chinese via Mixed Integer Optimization

Xiaoxi Luo, Weiwei Sun

This paper is concerned with phonetic reconstruction of the consonant system of Middle Chinese. We propose to cast the problem as a Mixed Integer Programming problem, which is able to automatically explore homophonic information from ancient rhyme dictionaries and phonetic information from modern Chinese dialects, the descendants of Middle Chinese. Numerical evaluation on a wide range of synthetic and real data demonstrates the effectiveness and robustness of the new method. We apply the method to information from Gu\v angy\` un (Author's note: The Gu\v angy\` un here is Chinese Pinyin in Latex, for HTML code, it should be Gu\ǎngy\ùn) and 20 modern Chinese dialects to obtain a new phonetic reconstruction result. A linguistically-motivated discussion of this result is also provided.

TabVer: Tabular Fact Verification with Natural Logic

Rami Aly, Andreas Vlachos

Final paper abstract: Fact verification on tabular evidence incentivises the use of symbolic reasoning models where a logical form is constructed (e.g. a LISP-style program), providing greater verifiability than fully neural approaches. However, these logical forms typically rely on well-formed tables, restricting their use in many scenarios. An emerging symbolic reasoning paradigm for textual evidence focuses on natural logic inference, which constructs proofs by modelling set-theoretic relations between a claim and its evidence in natural language. This approach provides flexibility and transparency but is less compatible with tabular evidence since the relations do not extend to arithmetic functions. We propose a set-theoretic interpretation of numerals and arithmetic functions in the context of natural logic, enabling the integration of arithmetic expressions in deterministic proofs. We leverage large language models to generate arithmetic expressions by generating questions about salient parts of a claim which are answered by executing appropriate functions on tables. In a few-shot setting on FEVEROUS, we achieve an accuracy of 71.4, outperforming both fully neural and symbolic reasoning models by 3.4 points. When evaluated on TabFact without any further training, our method remains competitive with an accuracy lead of 0.5 points.

LLM Reading Tea Leaves: Automatically Evaluating Topic Models with Large Language Models

Xiaohao Yang, He Zhao, Dinh Phung, Wray Buntine, Lan Du

Topic modeling has been a widely used tool for unsupervised text analysis. However, comprehensive evaluations of a topic model remain challenging. Existing evaluation methods are either less comparable across different models (e.g., perplexity) or focus on only one specific aspect of a model (e.g., topic quality or document representation quality) at a time, which is insufficient to reflect the overall model performance. In this paper, we propose WALM (Word Agreement with Language Model), a new evaluation method for topic modeling that considers the semantic quality of document representations and topics in a joint manner, leveraging the power of Large Language Models (LLMs). With extensive experiments involving different types of topic models, WALM is shown to align with human judgment and can serve as a complementary evaluation method to the existing ones, bringing a new perspective to topic modeling. Our software package is available at https://github.com/Xiaohao-Yang/Topic_Model_Evaluation.

Salute the Classic: Revisiting Challenges of Machine Translation in the Age of Large Language Models

Jianhui Pang, Fanghua Ye, Derek Fai Wong, Dian Yu, Shuming Shi, Zhaopeng Tu, Longyue Wang,

The evolution of Neural Machine Translation (NMT) has been significantly influenced by six core challenges, which have acted as benchmarks for progress in this field. This study revisits these challenges, offering insights into their ongoing relevance in the context of advanced Large Language Models (LLMs): domain mismatch, amount of parallel data, rare word prediction, translation of long sentences, attention model as word alignment}, and sub-optimal beam search}. Our empirical findings show that LLMs effectively reduce reliance on parallel data for major languages during pretraining and significantly improve translation of long sentences containing approximately 80 words, even translating documents up to 512 words. Despite these improvements, challenges in domain mismatch and rare word prediction persist. While word alignment and beam search challenges, specific to NMT, may not apply to LLMs, we identify three new challenges for LLM translation tasks: inference efficiency, translation of low-resource languages during pretraining, and human-aligned evaluation.

DEAR: Disentangled Event-Agnostic Representation Learning for Early Fake News Detection

Xiao Pu, Hao Wu, Xiuli Bi, Yu Wu, Xinbo Gao

Detecting fake news early is challenging due to the absence of labeled articles for emerging events in training data. To address this, we propose a Disentangled Event-Agnostic Representation (DEAR) learning approach. Our method begins with a BERT-based adaptive multi-grained semantic encoder that captures hierarchical and comprehensive textual representations of the input news content. To effectively separate latent authenticity-related and event-specific knowledge within the news content, we employ a disentanglement architecture. To further enhance the decoupling effect, we introduce a cross-perturbation mechanism that perturbs authenticity-related representation with the event-specific one, and vice versa, deriving a robust and discerning authenticity-related signal. Additionally, we implement a refinement learning scheme to minimize potential interactions between two decoupled representations, ensuring that the authenticity signal remains strong and unaffected by event-specific details. Experimental results demonstrate that our approach effectively mitigates the impact of event-specific influence, outperforming state-of-the-art methods. Particularly, it achieves a 6.0% improvement in accuracy on the PHEME dataset over MDDA, a similar approach that decouples latent content and style knowledge, in scenarios involving articles from unseen events different from the topics of the training set.

The Thai Universal Dependency Treebank

Panyut Sriwirote, Wei Qi Leong, Charin Polpanumas, Santhawat Thanyawong, William Tjhi, Wirote Aroonmanakun, Attapol T Rutherford

Automatic dependency parsing of Thai sentences has been underexplored, as evidenced by the lack of large Thai dependency treebanks with complete dependency structures and the lack of a published evaluation of state-of-the-art models, especially transformer-based parsers. In this work, we addressed these gaps by introducing the Thai Universal Dependency Treebank (TUD), a new Thai treebank consisting of 3,627 trees annotated according to the Universal Dependencies (UD) framework. We then benchmarked 92 dependency parsing models that incorporate pretrained transformers on Thai-PUD and our TUD, achieving state-of-the-art results and shedding light on the optimal model components for Thai dependency parsing. Our error analysis of the models also reveals that polyfunctional words, serial verb construction, and lack of rich morphosyntactic features present main challenges for Thai dependency parsing.

Investigating Adversarial Trigger Transfer in Large Language Models

Nicholas Carl Meade, Arkil Patel, Siva Reddy

Recent work has developed optimization procedures to find token sequences, called adversarial triggers, which can elicit unsafe responses from aligned language models. These triggers are believed to be highly transferable, i.e., a trigger optimized on one model can jailbreak other models. In this paper, we concretely show that such adversarial triggers are not consistently transferable. We extensively investigate trigger transfer amongst 13 open models and observe poor and inconsistent transfer. Our experiments further reveal a significant difference in robustness to adversarial triggers between models Aligned by Preference Optimization (APO) and models Aligned by Fine-Tuning (AFT). We find that APO models are extremely hard to jailbreak even when the trigger is optimized directly on the model. On the other hand, while AFT models may appear safe on the surface, exhibiting refusals to a range of unsafe instructions, we show that they are highly susceptible to adversarial triggers. Lastly, we observe that most triggers optimized on AFT models also generalize to new unsafe instructions from five diverse domains, further emphasizing their vulnerability. Overall, our work highlights the need for more comprehensive safety evaluations for aligned language models.

Few-shot Multilingual Open-domain QA from 5 Examples

Fan Jiang, Tom Drummond, Trevor Cohn

Recent approaches to multilingual open-domain question answering (MLODQA) have achieved promising results given abundant language-specific training data. However, the considerable annotation cost limits the probability of these methods for underrepresented languages. We introduce a \emph{few-shot} learning approach that generates large-scale multilingual data from large language models (LLMs) with minimal supervision for MLODQA. Our method begins with large-scale self-supervised pre-training by exploiting WikiData, followed by training on high-quality synthetic multilingual data generated by prompting LLMs with few-shot examples. The resulting model, FsModQA, significantly outperforms existing few-shot and supervised baselines in MLODQA and cross-lingual and monolingual retrieval. We further show our method can be extended for effective zero-shot adaptation to new languages through a \emph{cross-lingual prompting} strategy with only English-supervised data, making it a general and applicable solution for MLODQA tasks without costly large-scale annotation.

How "Real" is Your Real-Time Simultaneous Speech-to-Text Translation System?

Sara Papi, Peter Polak, Dominik Macháček, Ondřej Bojar

Simultaneous speech-to-text translation (SimulST) translates source-language speech into target-language text concurrently with the speaker's speech, ensuring low latency for better user comprehension. Despite its intended application to unbounded speech, most research has focused on human pre-segmented speech, simplifying the task and overlooking significant challenges. This narrow focus, coupled with widespread terminological inconsistencies affecting the task, is limiting the applicability of research outcomes to real-world applications, ultimately hindering progress in the field. Our extensive literature review of 110 papers not only reveals these critical issues in current research but also serves as the foundation for our key contributions, we: 1) define the steps and core components of a SimulST system, proposing a standardized terminology and taxonomy, 2) conduct a thorough analysis of community trends, and 3) offer concrete recommendations and future directions to bridge the gaps in existing literature, from evaluation frameworks to system architectures, for advancing the field towards more realistic and effective SimulST solutions.

Transformers as Transducers

Lena Strobl, Dana Angluin, David Chiang, Jonathan Rawski, Ashish Sabharwal

We study the sequence-to-sequence mapping capacity of transformers by relating them to finite transducers, and find that they can express surprisingly large classes of (total functional) transductions. We do so using variants of RASP, a programming language designed to help people "think like transformers," as an intermediate representation. We extend the existing Boolean variant B-RASP to sequence-to-sequence transductions and show that it computes exactly the first-order rational transductions (such as string rotation). Then, we introduce two new extensions. B-RASP[pos] enables calculations on positions (such as copying the first half of a string) and contains all first-order regular transductions. S-RASP adds prefix sum, which enables additional arithmetic operations (such as squaring a string) and contains all first-order polyregular transductions. Finally, we show that masked average-hard attention transformers can simulate S-RASP.

TaxoPro: A Plug-In LoRA-based Cross-Domain Method for Low-Resource Taxonomy Completion

Hongyuan Xu, Yuhang Niu, Ciyi Liu, Yanlong Wen, Xiaojie Yuan

Low-resource taxonomy completion aims to automatically insert new concepts into the existing taxonomy, in which only a few in-domain training samples are available. Recent studies have achieved considerable progress by incorporating prior knowledge from pre-trained language models (PLMs). However, these studies tend to overly rely on such knowledge and neglect the shareable knowledge across different taxonomies. In this paper, we propose TaxoPro, a plug-in LoRA-based cross-domain method, that captures shareable knowledge from the high-resource taxonomy to improve PLM-based low-resource taxonomy completion techniques. To prevent negative interference between domain-specific and domain-shared knowledge, TaxoPro decomposes cross-domain knowledge into domain-shared and domain-specific components, storing them using low-rank matrices (LoRA). Additionally, TaxoPro employs two auxiliary losses to regulate the flow of shareable knowledge. Experimental results demonstrate that TaxoPro improves PLM-based techniques, achieving state-of-the-art performance in completing low-resource taxonomies. Codes are available at https://github.com/cyclexu/TaxoPro.

Virtual Presentation

Session 2: V-Presentations (Monday 28 July, 11:00-12:30,)

SpatialWebAgent: Leveraging Large Language Models for Automated Spatial Information Extraction and Map Grounding

Shunfeng Zheng, Meng Fang, Ling Chen

Extracting spatial information from text is crucial for applications in geographic information systems (GIS), smart cities, disaster prevention, and logistics planning, helping decision-makers better understand geographic distributions and trends. However, the complex and unstructured nature of geographic expressions in text—including variations in place names, vague directions, and implicit spatial relationships—poses significant challenges for traditional methods. To overcome these challenges, we introduce SpatialWebAgent, an automated agent system that employs large language models (LLMs) to extract, standardize, and ground spatial information from natural language onto maps. SpatialWebAgent addresses the issues posed by diverse and ambiguous geographic expressions—ranging from varying place names and vague directional cues to implicit spatial relationships that require flexible combinations of localization functions—by leveraging the deep semantic understanding of LLMs. Our system extracts different types of spatial entities, then applies a series of geospatial functions to convert them into precise coordinates via geocoding APIs, and finally visualizes the results on interactive maps. Our demo is available at: https://sites.google.com/view/SpatialWebAgent.

ROGRAG: A Robustly Optimized GraphRAG Framework

Zhefan Wang, Huanjun Kong, Jie Ying, Wanli Ouyang, Nanqing Dong

Large language models (LLMs) commonly struggle with specialized or emerging topics which are rarely seen in the training corpus. Graph-based retrieval-augmented generation (GraphRAG) addresses this by structuring domain knowledge as a graph for dynamic retrieval. However, existing pipelines involve complex engineering workflows, making it difficult to isolate the impact of individual components. It is also challenging to evaluate the retrieval effectiveness due to the overlap between the pretraining and evaluation datasets. In this work, we introduce ROGRAG, a Robustly Optimized GraphRAG framework. Specifically, we propose a multi-stage retrieval mechanism that integrates dual-level with logic form methods to improve retrieval robustness without increasing computational cost. To further refine the system, we incorporate various result verification methods and adopt an incremental database construction approach. Through extensive ablation experiments, we rigorously assess the effectiveness of each component. Our implementation includes comparative experiments on SeedBench, where Qwen2.5-7B-Instruct initially underperformed. ROGRAG significantly improves the score from 60.0% to 75.0% and outperforms mainstream methods. Experiments on domain-specific datasets reveal that dual-level retrieval enhances fuzzy matching, while logic form retrieval improves structured reasoning, highlighting the importance of adaptive retrieval. ROGRAG is deployed on an online research platform https://seedllm.org.cn and released as an open-source resource https://github.com/tpoisonooo/ROGRAG. It supports easy installation with pip and a demo video https://www.youtube.com/watch?v=NL1_cuIiX-w is provided.

MIND: A Multi-agent Framework for Zero-shot Harmful Meme Detection

Ziyan Liu, Chunxiao Fan, Haoran Lou, Yuexin Wu, Kaiwei Deng

The rapid expansion of memes on social media has highlighted the urgent need for effective approaches to detect harmful content. However, traditional data-driven approaches struggle to detect new memes due to their evolving nature and the lack of up-to-date annotated data. To address this issue, we propose MIND, a multi-agent framework for zero-shot harmful meme detection that does not rely on annotated data. MIND implements three key strategies: 1) We retrieve similar memes from an unannotated reference set to provide contextual information. 2) We propose a bi-directional insight derivation mechanism to extract a comprehensive understanding of similar memes. 3) We then employ a multi-agent debate mechanism to ensure robust decision-making through reasoned arbitration. Extensive experiments on three meme datasets demonstrate that our proposed framework not only outperforms existing zero-shot approaches but also shows strong generalization across different model architectures and parameter scales, providing a scalable solution for harmful meme detection.

ProcessBench: Identifying Process Errors in Mathematical Reasoning

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, Junyang Lin

As language models regularly make mistakes when solving math problems, automated identification of errors in the reasoning process becomes increasingly significant for their scalable oversight. In this paper, we introduce ProcessBench for measuring the ability to identify erroneous steps in mathematical reasoning. It consists of 3,400 test cases, primarily focused on competition- and Olympiad-level math problems. Each test case contains a step-by-step solution with error location annotated by human experts. Models are required to identify the earliest step that contains an error, or conclude that all steps are correct. We conduct extensive evaluation on ProcessBench, involving two types of models: process reward models (PRMs) and critic models, where for the latter we prompt general language models to critique each solution step by step. We draw two main observations: (1) Existing PRMs typically fail to generalize to more challenging math problems beyond GSM8K and MATH. They underperform both critic models (i.e., prompted general language models) and our own trained PRM that is straightforwardly fine-tuned on the PRM800K dataset. (2) The best open-source model, QwQ-32B-Preview, has demonstrated the critique capability competitive with the proprietary model GPT-4o, despite that it still lags behind the reasoning-specialized o1-mini. We hope ProcessBench can foster future research in reasoning process assessment, paving the way toward scalable oversight of language models.

Model Extrapolation Expedites Alignment

Chujie Zheng, Ziqi Wang, Heng Ji, Minlie Huang, Nanyun Peng

Given the high computational cost of preference alignment training of large language models (LLMs), exploring efficient methods to reduce the training overhead remains an important and compelling research problem. Motivated by the observation that alignment training typically involves only small parameter changes without injecting new knowledge into models, we propose a straightforward method called ExPO (model extrapolation) to expedite LLMs' alignment with human preferences. Given a partially-trained model and its initial SFT checkpoint, ExPO improves the implicit optimization objective of alignment training by simply amplifying the parameter change based on a first-order approximation, without any additional training overhead. Through controlled experiments, we demonstrate that ExPO boosts a DPO model trained with only 20% steps to outperform the fully-trained one. Moreover, we show that ExPO notably improves existing open-source LLMs (ranging from 1.8B to 70B parameters) on the leading AlpacaEval 2.0 and MT-Bench benchmarks, which highlights ExPO's broader utility in efficiently enhancing LLM alignment.

MPVStance: Mitigating Hallucinations in Stance Detection with Multi-Perspective Verification

ZhaoDan Zhang, Zhao Zhang, Jin Zhang, Hui Xu, Xueqi Cheng

Stance detection is a pivotal task in Natural Language Processing (NLP), identifying textual attitudes toward various targets. Despite advances in using Large Language Models (LLMs), challenges persist due to hallucination-models generating plausible yet inaccurate content. Addressing these challenges, we introduce MPVStance, a framework that incorporates Multi-Perspective Verification (MPV) with Retrieval-Augmented Generation (RAG) across a structured five-step verification process. Our method enhances stance detection by rigorously validating each response from factual accuracy, logical consistency, contextual relevance, and other perspectives. Extensive testing on the SemEval-2016 and VAST datasets, including scenarios that challenge existing methods and comprehensive ablation studies, demonstrates that MPVStance significantly outperforms current models. It effectively mitigates hallucination issues and sets new benchmarks for reliability and accuracy in stance detection, particularly in zero-shot, few-shot, and challenging scenarios.

Exploring Forgetting in Large Language Model Pre-Training

Chonghua Liao, Ruobing Xie, Xingwu Sun, Haowen Sun, Zhanhui Kang

Catastrophic forgetting remains a formidable obstacle to building an omniscient model in large language models (LLMs). Despite the pioneering research on task-level forgetting in LLM fine-tuning, there is scant focus on forgetting during pre-training. We systematically explored the existence and measurement of forgetting in pre-training, questioning traditional metrics such as perplexity (PPL) and introducing new metrics to better detect entity memory retention. Based on our revised assessment of forgetting metrics, we explored low-cost, straightforward methods to mitigate forgetting during the pre-training phase. In addition, we carefully analyzed the learning curves, offering insights into the dynamics of forgetting. Extensive evaluations and analyses on forgetting of pre-training could facilitate future research on LLMs.

Iron Sharpens Iron: Defending Against Attacks in Machine-Generated Text Detection with Adversarial Training

Yuanfan Li, Zhaohan Zhang, Chengzhengxu Li, Chao Shen, Xiaoming Liu

Machine-generated Text (MGT) detection is crucial for regulating and attributing online texts. While the existing MGT detectors achieve strong performance, they remain vulnerable to simple perturbations and adversarial attacks. To build an effective defense against malicious perturbations, we view MGT detection from a threat modeling perspective, that is, analyzing the model's vulnerability from an adversary's point of view and exploring effective mitigations. To this end, we introduce an adversarial framework for training a robust MGT detector, named GREedy Adversary PromoTed DefendER (GREATER). The GREATER consists of two key components: an adversary GREATER-A and a detector GREATER-D. The GREATER-D learns to defend against the adversarial attack from GREATER-A and generalizes the defense to other attacks. GREATER-A identifies and perturbs the critical tokens in embedding space, along with greedy search and pruning to generate stealthy and disruptive adversarial examples. Besides, we update the GREATER-A and GREATER-D synchronously, encouraging the GREATER-D to generalize its defense to different attacks and varying attack intensities. Our experimental results across 10 text perturbation strategies and 6 adversarial attacks show that our GREATER-D reduces the Attack Success Rate (ASR) by 0.67% compared with SOTA defense methods while our GREATER-A is demonstrated to be more effective and efficient than SOTA attack approaches. Codes and dataset are available in https://github.com/Liyuuuu111/GREATER.

Taming LLMs with Gradient Grouping

Siyuan Li, Juanxi Tian, Zedong Wang, Xin Jin, Zicheng Liu, Wentao Zhang, Dan Xu

Training large language models (LLMs) poses challenges due to their massive scale and heterogeneous architectures. While adaptive optimizers like AdamW help address gradient variations, they still struggle with efficient and effective parameter-wise learning rate estimation, resulting in training instability, slow convergence, and poor compatibility with parameter-efficient fine-tuning (PEFT) techniques. This work introduces Scaling with Gradient Grouping (SGG), an optimizer wrapper that improves adaptive learning rate estimation by dynamic grouping and group-specific scaling. SGG first groups gradient statistics in each layer into clusters and then applies cluster-specific scaling to calibrate learning rates for each parameter, thus imposing collective group-wise constraints while maintaining precise per-parameter adaptation. Experiments on diverse (M)LLM benchmarks show that SGG integrates seamlessly with existing optimizers, and offers consistent gains and faster convergence over baselines, with various model sizes. Its stability across varying batch sizes and learning rates establishes SGG as a robust choice for LLM optimization.

Retrieve to Explain: Evidence-driven Predictions for Explainable Drug Target Identification

Ravi Patel, Angus Brayne, Rogier Hintzen, Daniel Jaroslawicz, Georgiana Neculae, Dane S. Corneil

Language models hold incredible promise for enabling scientific discovery by synthesizing massive research corpora. Many complex scientific research questions have multiple plausible answers, each supported by evidence of varying strength. However, existing language models lack the capability to quantitatively and faithfully compare answer plausibility in terms of supporting evidence. To address this, we introduce Retrieve to Explain (R2E), a retrieval-based model that scores and ranks all possible answers to a research question based on evidence retrieved from a document corpus. The architecture represents each answer only in terms of its supporting evidence, with the answer itself masked. This allows us to extend feature attribution methods such as Shapley values, to transparently attribute answer scores to supporting evidence at inference time. The architecture also allows incorporation of new evidence without retraining, including non-textual data modalities templated into natural language. We developed R2E for the challenging scientific discovery task of drug target identification, a human-in-the-loop process where failures are extremely costly and explainability paramount. When predicting whether drug targets will subsequently be confirmed as efficacious in clinical trials, R2E not only matches non-explainable literature-based models but also surpasses a genetics-based target identification approach used throughout the pharmaceutical industry.

TEACH: A Contrastive Knowledge Adaptive Distillation Framework for Ancient Chinese Understanding

Yuting Wei, Qi Meng, Yuanxing Xu, Bin Wu

Traditional methods for processing classical Chinese typically segment language understanding into discrete tasks, which overlook crucial background information and reduce user engagement. Large language models (LLMs) provide integrated solutions, yet they entail high computational costs and risks of generating inaccurate historical information. To tackle these challenges, we propose a novel framework, TEACH (conTrastive knowlEdge Adaptive distillation with enhanCed Historical interpretability), which focuses on classical Chinese understanding by integrating word sense disambiguation with sentence translation. This integration leverages a confidence-annotated knowledge base and a step-by-step Chain-of-Thought prompting mechanism to minimize hallucinations and improve semantic analysis. Moreover, TEACH employs contrastive distillation learning to efficiently transfer capabilities from larger models to smaller ones (e.g., Qwen2-1.5B), addressing overly liberal translations. Additionally, we introduce an innovative generation evaluation metric using iterative word alignment, enhancing LLM performance assessments by distinguishing additional information and addressing excessive translation issues. Experiments conducted on real-world datasets validate TEACH's efficacy in classical Chinese educational scenarios.

Confidence v.s. Critique: A Decomposition of Self-Correction Capability for LLMs

Zhe Yang, Yichang Zhang, Yudong Wang, Ziyao Xu, Junyang Lin, Zhifang Sui

Large Language Models (LLMs) can correct their self-generated responses, but a decline in accuracy after self-correction is also witnessed. To have a deeper understanding of self-correction, we endeavor to decompose, evaluate, and analyze the self-correction behaviors of LLMs. By enumerating and analyzing answer correctness before and after self-correction, we decompose the self-correction capability into confidence (being confident to correct answers) and critique (turning wrong answers to correct) capabilities, and propose two metrics from a probabilistic perspective to measure these 2 capabilities, along with another metric for overall self-correction capability evaluation. Based on our decomposition and evaluation metrics, we conduct extensive experiments and draw some empirical conclusions. For example, we find different models can exhibit distinct behaviors: some models are confident while others are more critical. We also find the trade-off between the two capabilities (i.e. improving one can lead to a decline in the other) when manipulating model self-correction behavior by prompts or in-context learning. Further, we find a simple yet efficient strategy to improve self-correction capability by transforming Supervision Fine-Tuning (SFT) data format, and our strategy outperforms vanilla SFT in both capabilities and achieves much higher accuracy after self-correction.

ProtoLens: Advancing Prototype Learning for Fine-Grained Interpretability in Text Classification

Bowen Wei, Ziwei Zhu

In this work, we propose ProtoLens, a novel prototype-based model that provides fine-grained, sub-sentence level interpretability for text classification. ProtoLens uses a Prototype-aware Span Extraction module to identify relevant text spans associated with learned prototypes and a Prototype Alignment mechanism to ensure prototypes are semantically meaningful throughout training. By aligning the prototype embeddings with human-understandable examples, ProtoLens provides interpretable predictions while maintaining competitive accuracy. Extensive experiments demonstrate that ProtoLens outperforms both prototype-based and non-interpretable baselines on multiple text classification benchmarks. Code and data are available at .

AnRe: Analogical Replay for Temporal Knowledge Graph Forecasting

Guo Tang, Zheng Chu, Wenxiang Zheng, Junjia Xiang, Yizhuo Li, Weihao Zhang, Ming Liu, Bing Qin

Temporal Knowledge Graphs (TKGs) are vital for event prediction, yet current methods face limitations. Graph neural networks mainly depend on structural information, often overlooking semantic understanding and requiring high computational costs. Meanwhile, Large Language Models (LLMs) support zero-shot reasoning but lack sufficient capabilities to grasp the laws of historical event development. To tackle these challenges, we introduce a training-free Analogical Replay (AnRe) reasoning framework. Our approach retrieves similar events for queries through semantic-driven clustering and builds comprehensive historical contexts using a dual history extraction module that integrates long-term and short-term history. It then uses LLMs to generate analogical reasoning examples as contextual inputs, enabling the model to deeply understand historical patterns of similar events and improve its ability to predict unknown ones. Our experiments on four benchmarks show that AnRe significantly exceeds traditional training and existing LLM-based methods. Further ablation studies also confirm the effectiveness of the dual history extraction and analogical replay mechanisms.

MoC: Mixtures of Text Chunking Learners for Retrieval-Augmented Generation System

Jihao Zhao, Zhiyuan Ji, Zhaoxin Fan, Hanyu Wang, Simin Niu, Bo Tang, Feiyu Xiong, Zhiyu li

Retrieval-Augmented Generation (RAG), while serving as a viable complement to large language models (LLMs), often overlooks the crucial aspect of text chunking within its pipeline. This paper initially introduces a dual-metric evaluation method, comprising Boundary Clarity and Chunk Stickiness, to enable the direct quantification of chunking quality. Leveraging this assessment method, we highlight the inherent limitations of traditional and semantic chunking in handling complex contextual nuances, thereby substantiating the necessity of integrating LLMs into chunking process. To address the inherent trade-off between computational efficiency and chunking precision in LLM-based approaches, we devise the granularity-aware Mixture-of-Chunkers (MoC) framework, which consists of a three-stage processing mechanism. Notably, our objective is to guide the chunker towards generating a structured list of chunking regular expressions, which are subsequently employed to extract chunks from the original text. Extensive experiments demonstrate that both our proposed metrics and the MoC framework effectively settle challenges of the chunking task, revealing the chunking kernel while enhancing the performance of the RAG system.

Unlocking General Long Chain-of-Thought Reasoning Capabilities of Large Language Models via Representation Engineering

Xinyu Tang, Xiaolei Wang, Zhihao Lv, Yingqian Min, Xin Zhao, Binbin Hu, Ziqi Liu, Zhiqiang Zhang

Recent advancements in long chain-of-thoughts (long CoTs) have significantly improved the reasoning capabilities of large language models (LLMs). Existing work finds that the capability of long CoT reasoning can be efficiently elicited by tuning on only a few examples and can easily transfer to other tasks. This motivates us to investigate whether long CoT reasoning is a general capability for LLMs. In this work, we conduct an empirical analysis for this question from the perspective of representation. We find that LLMs do encode long CoT reasoning as a general capability, with a clear distinction from vanilla CoTs. Furthermore, domain-specific representations are also required for the effective transfer of long CoT reasoning. Inspired by these findings, we propose GLORE, a novel representation engineering method to unleash the general long CoT reasoning capabilities of LLMs. Extensive experiments demonstrate the effectiveness and efficiency of GLORE in both in-domain and cross-domain scenarios. The code is available at https://github.com/txy77/GLoRE.

PIC: Unlocking Long-Form Text Generation Capabilities of Large Language Models via Position ID Compression

Haoran Que, Wenge Rong

Long-context understanding is crucial for large language models (LLMs) and has become a fundamental capability for most LLMs. However, beyond the focus on "input-long", the ability to "output-long" is equally significant, yet it remains underexplored. To address this limitation, we propose a simple, efficient, and plug-in approach, Position ID Compression (PIC), to unlock the long-form text generation potential of LLMs. The idea is straightforward: by compressing the position ids of the context, we provoke and guide LLMs to generate coherent and longer output. Specifically, we find that directly reducing the position ids by a fixed ratio significantly impacts the generation quality. To mitigate this, we propose two variants of PIC: NTK-aware PIC and Dynamic PIC. Without additional training, both methods enable LLMs to extend their generation length by approximately 1.5 times without compromising generation quality. Furthermore, by integrating supervised fine-tuning (SFT) with PIC, we propose PIC-SFT, which further improves LLMs' long-form text generation capabilities, achieving top performance on HelloBench and LongBench-Write. Extensive experiments demonstrate the effectiveness of our approach.

AXIS: Efficient Human-Agent-Computer Interaction with API-First LLM-Based Agents

Junting Lu, Zhiyang Zhang, Fangkai Yang, Jue Zhang, Lu Wang, Chao Du, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, Qi Zhang

Multimodal large language models (MLLMs) have enabled LLM-based agents to directly interact with application user interfaces (UIs), enhancing agents' performance in complex tasks. However, these agents often suffer from high latency and low reliability due to the extensive sequential UI interactions. To address this issue, we propose AXIS, a novel LLM-based agents framework that prioritize actions through application programming interfaces (APIs) over UI actions. This framework also facilitates the creation and expansion of APIs through automated exploration of applications. Our experiments on Microsoft Word demonstrate that AXIS reduces task completion time by 65%-70% and cognitive workload by 38%-53%, while maintaining accuracy of 97%-98% compared to humans. Our work contributes to a new human-agent-computer interaction (HACI) framework and explores a fresh UI design principle for application providers to turn applications into agents in the era of LLMs, paving the way towards an agent-centric operating system (Agent OS). The code and dataset will be available at https://aka.ms/haci_axis.

Watermarking Large Language Models: An Unbiased and Low-risk Method

Minjia Mao, Dongjun Wei, Zeyu Chen, Xiao Fang, Michael Chau

Recent advancements in large language models (LLMs) have highlighted the risk of misusing them, raising the need for accurate detection of LLM-generated content. In response, a viable solution is to inject imperceptible identifiers into LLMs, known as watermarks. Our research extends the existing watermarking methods by proposing the novel Sampling One Then Accepting (STA-1) method. STA-1 is an unbiased watermark that preserves the original token distribution in expectation and has a lower risk of producing unsatisfactory outputs in low-entropy scenarios compared to existing unbiased watermarks. In watermark detection, STA-1 does not require prompts or a white-box LLM, provides statistical guarantees, demonstrates high efficiency in detection time, and remains robust against various watermarking attacks. Experimental results on low-entropy and high-entropy datasets demonstrate that STA-1 achieves the above properties simultaneously, making it a desirable solution for watermarking LLMs. Implementation codes for this study are available online.

SEUF: Is Unlearning One Expert Enough for Mixture-of-Experts LLMs?

Haomin Zhuang, Yihua Zhang, Kehan Guo, Jinghan Jia, Gaowen Liu, Sijia Liu, Xiangliang Zhang

Recent advancements in LLMs unlearning have shown remarkable success in removing unwanted data-model influences while preserving the model's utility for legitimate knowledge. Despite these strides, sparse Mixture-of-Experts (MoE) LLMs--a key subset of the LLM family--have remained unexplored in the context of unlearning. As MoE LLMs are celebrated for their exceptional performance, we ask:How can unlearning be performed effectively and efficiently on MoE LLMs? Our pilot study shows that the dynamic routing nature of MoE LLMs introduces unique challenges, leading to excessive forgetting, uncontrolled knowledge erasure and substantial utility drops when existing unlearning methods are applied. To address this, we propose a novel Selected-Expert Unlearning Framework (SEUF). Through expert attribution, unlearning is concentrated on the most actively engaged experts for the specified knowledge. Concurrently, an anchor loss is applied to the router to stabilize the active state of this targeted expert, ensuring focused and controlled unlearning. SEUF is compatible with various standard unlearning algorithms. Extensive experiments demonstrate that SEUF enhances both forget quality up to 5% and model utility by 35% on MoE LLMs across various benchmarks and LLM architectures (compared to standard unlearning algorithms), while only unlearning 0.06% of the model parameters.

SkillVerse : Assessing and Enhancing LLMs with Tree Evaluation

Yufei Tian, Jiao Sun, Nanyun Peng, Zizhao Zhang

As language models evolve to tackle complex, multifaceted tasks, their evaluation must adapt to capture this intricacy. A granular, skill-specific understanding of model capabilities can empower researchers to make informed model development plans. In this paper, we introduce SkillVerse, an unsupervised tree-structured diagnosis framework for understanding model proficiency in specific abilities. With LLM as a judge, SkillVerse first critiques the model responses, and then organizes them into a hierarchical structure termed dendrogram. Given proficiency at arbitrary levels of granularity, SkillVerse is flexible to produce insights of behaviors of modern large models. We also demonstrate its efficacy in two downstream tasks: 1) improving model in-context learning by 25% using a tree-search algorithm to select more informative few-shot demonstrations, and 2) accurately predicting new model weaknesses with a 55% success rate, 22% higher than without SkillVerse.

SelfElicit: Your Language Model Secretly Knows Where is the Relevant Evidence

Zhining Liu, Rana Ali Amjad, Ravinarayana Adkathimar, Tianxin Wei, Hanghang Tong

Providing Language Models (LMs) with relevant evidence in the context (either via retrieval or user-provided) can significantly improve their ability to provide better-grounded responses. However, recent studies have found that LMs often struggle to fully comprehend and utilize key evidence from the context, especially when it contains noise and irrelevant information—an issue common in real-world scenarios. To address this, we propose SelfElicit, an inference-time approach that helps LMs focus on key contextual evidence through self-guided explicit highlighting. By leveraging the inherent evidence-finding capabilities of LMs using the attention scores of deeper layers, our method automatically identifies and emphasizes key evidence within the input context, facilitating more accurate and grounded responses without additional training or iterative prompting. We demonstrate that SelfElicit brings consistent and significant improvement on multiple evidence-based QA tasks for various LM families while maintaining computational efficiency. Our code and documentation are available at https://github.com/ZhiningLiu1998/SelfElicit.

The Male CEO and the Female Assistant: Evaluation and Mitigation of Gender Biases in Text-To-Image Generation of Dual Subjects

Yixin Wan, Kai-Wei Chang

Recent large-scale T2I models like DALLE-3 have made progress in reducing gender stereotypes when generating single-person images. However, significant biases remain when generating images with more than one person. To systematically evaluate this, we propose the **Paired Stereotype Test (PST)** framework, which queries T2I models to depict two individuals assigned with male-stereotyped and female-stereotyped social identities, respectively (e.g. "a CEO" and "an Assistant"). This contrastive setting often triggers T2I models to generate gender-stereotyped images. Using PST, we evaluate two aspects of gender biases -- the well-known **bias in gendered occupation** and a novel aspect: **bias in organizational power**. Experiments show that **over 74% images generated by DALLE-3 display gender-occupational biases**. Additionally, compared to single-person settings, DALLE-3 is more likely to perpetuate male-associated stereotypes under PST. We further propose **FairCritic**, a novel and interpretable framework that leverages an LLM-based critic model to i) detect bias in generated images, and ii) adaptively provide feedback to T2I models for improving fairness. FairCritic achieves near-perfect fairness on PST, overcoming the limitations of previous prompt-based intervention approaches.

Identifying Cellular Niches in Spatial Transcriptomics: An Investigation into the Capabilities of Large Language Models

Huanhuan Wei, Xiao Luo, Hongyi Yu, Jinping Liang, Luning Yang, Lixing Lin, Alexandra Popa, Xiting Yan

Spatial transcriptomic technologies enable measuring gene expression profile and spatial information of cells in tissues simultaneously. Clustering of captured cells/spots in the spatial transcriptomic data is crucial for understanding tissue niches and uncovering disease-related changes. Current methods to cluster spatial transcriptomic data encounter obstacles, including inefficiency in handling multi-replicate data, lack of prior knowledge incorporation, and producing uninterpretable cluster labels. We introduce a novel approach, LLMiniST, to identify spatial niche using a zero-shot large language models (LLMs) by transforming spatial transcriptomic data into spatial context prompts, leveraging gene expression of neighboring cells/spots, cell type composition, tissue information, and external knowledge. The model was further enhanced using a two-stage fine-tuning strategy for improved generalizability. We also develop a user-friendly annotation tool to accelerate the creation of well-annotated spatial dataset for fine-tuning. Comprehensive method performance evaluations showed that both zero-shot and fine-tunned LLMiniST had superior performance than current non-LLM methods in many circumstances. Notably, the two-stage fine-tuning strategy facilitated substantial cross-subject generalizability. The results demonstrate the feasibility of LLMs for tissue niche identification using spatial transcriptomic data and the potential of LLMs as a scalable solution to efficiently integrate minimal human guidance for improved performance in large-scale datasets.

The Lawyer That Never Thinks: Consistency and Fairness as Keys to Reliable AI

Dana R Alsagheer, Abdulrahman Kamal, Mohammad Kamal, Cosmo Yang Wu, Weidong Shi

Large Language Models (LLMs) are increasingly used in high-stakes domains like law and research, yet their inconsistencies and response instability raise concerns about trustworthiness. This study evaluates six leading LLMs—GPT-3.5, GPT-4, Claude, Gemini, Mistral, and LLaMA 2—on rationality, stability, and ethical fairness through reasoning tests, legal challenges, and bias-sensitive scenarios. Results reveal significant inconsistencies, highlighting trade-offs between model scale, architecture, and logical coherence. These findings underscore the risks of deploying LLMs in legal and policy settings, emphasizing the need for AI systems that prioritize transparency, fairness, and ethical robustness.

Cracking Factual Knowledge: A Comprehensive Analysis of Degenerate Knowledge Neurons in Large Language Models

Yuheng Chen, Pengfei Cao, Yubo Chen, Yining Wang, Shengping Liu, Kang Liu, Jun Zhao

Knowledge neuron theory provides a key approach to understanding the mechanisms of factual knowledge in Large Language Models (LLMs), which suggests that facts are stored within multi-layer perceptron neurons. This paper further explores **Degenerate Knowledge Neurons** (DKNs), where distinct sets of neurons can store identical facts, but unlike simple redundancy, they also participate in storing other different facts. Despite the novelty and unique properties of this concept, it has not been rigorously defined and systematically studied. Our contributions are: (1) We pioneer the study of structures in knowledge neurons by analyzing weight connection patterns, providing a comprehensive definition of DKNs from both functional and structural aspects. (2) Based on this definition, we develop the **Neuronal Topology Clustering** method, leading to a more accurate DKN identification. (3) We demonstrate the practical applications of DKNs in two aspects: guiding LLMs to learn new knowledge and relating to LLMs' robustness against input errors.

Introducing Graph Context into Language Models through Parameter-Efficient Fine-Tuning for Lexical Relation Mining

Jingwen Sun, Zhiyi Tian, Yu He, Jingwei Sun, Guangzhong Sun

Lexical relation refers to the way words are related within a language. Prior work has demonstrated that pretrained language models (PLMs) can effectively mine lexical relations between word pairs. However, they overlook the potential of graph structures composed of lexical relations, which can be integrated with the semantic knowledge of PLMs. In this work, we propose a parameter-efficient fine-tuning method through graph context, which integrates graph features and semantic representations for lexical relation classification (LRC) and lexical entailment (LE) tasks. Our experiments show that graph features can help PLMs better understand more complex lexical relations, establishing a new state-of-the-art for LRC and LE. Finally, we perform an error analysis, identifying the bottlenecks of language models in lexical relation mining tasks and providing insights for future improvements.

Praetor: A Fine-Grained Generative LLM Evaluator with Instance-Level Customizable Evaluation Criteria

Yongqi Leng, Renren Jin, Yue chen, Zhuowen Han, Ling Shi, Jianxiang Peng, Lei Yang, Juesi Xiao, Deyi Xiong

With the increasing capability of large language models (LLMs), LLM-as-a-judge has emerged as a new evaluation paradigm. Compared with traditional automatic and manual evaluation, LLM evaluators exhibit better interpretability and efficiency. Despite this, existing LLM evaluators suffer from limited use scenarios and poor flexibility. To mitigate these issues, we propose Praetor, a fine-grained generative LLM evaluator with instance-level customazable evaluation criteria. To train Praetor, we curate a large-scale dataset guided with a hierarchical guideline covering a wide range of tasks and instance-level evaluation criteria. We train Praetor on this dataset in a multi-task learning fashion, which enables to evaluate LLMs in either pointwise grading or pairwise comparison way and support two languages simultaneously with a high flexibility of setting evaluation criteria. Extensive experiments demonstrate that Praetor outperforms previous LLM evaluators and instruction-tuned LLMs on multiple benchmarks, setting new SOTA results. It also exhibits the potential for generating critiques as scalable feedback to further improve LLMs. Our model and related resources are released at .

The Knowledge Microscope: Features as Better Analytical Lenses than Neurons

Yuheng Chen, Pengfei Cao, Kang Liu, Jun Zhao

We demonstrate that features, rather than neurons, serve as superior analytical units for understanding the mechanisms of factual knowledge in Language Models (LMs). Previous studies primarily utilize MLP neurons as units of analysis; however, neurons suffer from polysemanticity, leading to limited knowledge expression and poor interpretability. We first conduct preliminary experiments to validate that SAE can effectively decompose neurons into features. With this established, our core findings reveal three key advantages of features over neurons: (1) Features exhibit stronger influence on knowledge expression and superior interpretability. (2) Features demonstrate enhanced monosemanticity, showing distinct activation patterns between related and unrelated facts. (3) Feature-based method demonstrates superior performance over neuron-based approaches in erasing privacy-sensitive information from LMs. Additionally, we propose FeatureEdit, the first feature-based editing method. Code and dataset will be available.

Dynamic Parallel Tree Search for Efficient LLM Reasoning

Yifu Ding, Wentao Jiang, Shunyu Liu, Yongcheng Jing, Jinyang Guo, Yingjie Wang, Jing Zhang, Zengmao Wang, Ziwei Liu, Bo Du, Xianglong Liu, Dacheng Tao

Tree of Thoughts (ToT) enhances Large Language Model (LLM) reasoning by structuring problem-solving as a spanning tree. However, recent methods focus on search accuracy while overlooking computational efficiency. The challenges of accelerating the ToT lie in the frequent switching of reasoning focus, and the redundant exploration of suboptimal solutions. To alleviate this dilemma, we propose Dynamic Parallel Tree Search (DPTS), a novel parallelism framework that aims to dynamically optimize the reasoning path in inference. It includes the Parallelism Streamline in the generation phase to build up a flexible and adaptive parallelism with arbitrary paths by cache management and alignment. Meanwhile, the Search and Transition Mechanism filters potential candidates to dynamically maintain the reasoning focus on more possible solutions with less redundancy. Experiments on Qwen-2.5 and Llama-3 on math and code datasets show that DPTS significantly improves efficiency by 2-4\times on average while maintaining or even surpassing existing reasoning algorithms in accuracy, making ToT-based reasoning more scalable and computationally efficient. Codes are released at: https://github.com/yifu-ding/DPTS.

MiniLongBench: The Low-cost Long Context Understanding Benchmark for Large Language Models

Zhongzhan Huang, Guoming Ling, Shanshan Zhong, Hefeng Wu, Liang Lin

Long Context Understanding (LCU) is a critical area for exploration in current large language models (LLMs). However, due to the inherently lengthy nature of long-text data, existing LCU benchmarks for LLMs often result in prohibitively high evaluation costs, like testing time and inference expenses. Through extensive experimentation, we discover that existing LCU benchmarks exhibit significant redundancy, which means the inefficiency in evaluation. In this paper, we propose a concise data compression method tailored for long-text data with sparse information characteristics. By pruning the well-known LCU benchmark LongBench, we create MiniLongBench. This benchmark includes only 237 test samples across six major task categories and 21 distinct tasks. Through empirical analysis of over 60 LLMs, MiniLongBench achieves an average evaluation cost reduced to only 4.5% of the original while maintaining an average rank correlation coefficient of 0.97 with LongBench results. Therefore, our MiniLongBench, as a low-cost benchmark, holds great potential to substantially drive future research into the LCU capabilities of LLMs.

Divide-Then-Align: Honest Alignment based on the Knowledge Boundary of RAG

Xin Sun, Jianan Xie, Zhongqi Chen, Qiang Liu, Shu Wu, Yuehe Chen, Bowen Song, Zilei Wang, Weiqiang Wang, Liang Wang

Large language models (LLMs) augmented with retrieval systems have significantly advanced natural language processing tasks by integrating external knowledge sources, enabling more accurate and contextually rich responses. To improve the robustness of such systems against noisy retrievals, Retrieval-Augmented Fine-Tuning (RAFT) has emerged as a widely adopted method. However, RAFT conditions models to generate answers even in the absence of reliable knowledge. This behavior undermines their reliability in high-stakes domains, where acknowledging uncertainty is critical. To address this issue, we propose Divide-Then-Align (DTA), a post-training approach designed to endow RAG systems with the ability to respond with "I don't know" when the query is out of the knowledge boundary of both the retrieved passages and the model's internal knowledge. DTA divides data samples into four knowledge quadrants and constructs tailored preference data for each quadrant, resulting in a curated dataset for Direct Preference Optimization (DPO). Experimental results on three benchmark datasets demonstrate that \ourmethod effectively balances accuracy with appropriate abstention, enhancing the reliability and trustworthiness of retrieval-augmented systems.

TROVE: A Challenge for Fine-Grained Text Provenance via Source Sentence Tracing and Relationship Classification

Junnan Zhu, Min Xiao, Yining Wang, Feifei Zhai, Yu Zhou, Chengqing Zong

LLMs have achieved remarkable fluency and coherence in text generation, yet their widespread adoption has raised concerns about content reliability and accountability. In high-stakes domains, it is crucial to understand where and how the content is created. To address this, we introduce the Text pROVEnance (TROVE) challenge, designed to trace each sentence of a target text back to specific source sentences within potentially lengthy or multi-document inputs. Beyond identifying sources, TROVE annotates the fine-grained relationships (quotation, compression, inference, and others), providing a deep understanding of how each target sentence is formed. To benchmark TROVE, we construct our dataset by leveraging three public datasets covering 11 diverse scenarios (e.g., QA and summarization) in English and Chinese, spanning source texts of varying lengths (0–5k, 5–10k, 10k+), emphasizing the multi-document and long-document settings essential for provenance. To ensure high-quality data, we employ a three-stage annotation process: sentence retrieval, GPT-4o provenance, and human provenance. We evaluate 11 LLMs under direct prompting and retrieval-augmented paradigms, revealing that retrieval is essential for robust performance, larger models perform better in complex relationship classification, and closed-source models often lead, yet open-source models show significant promise, particularly with retrieval augmentation. We make our dataset available here: https://github.com/ZNLP/ZNLP-Dataset.

From Objectives to Questions: A Planning-based Framework for Educational Mathematical Question Generation

Cheng Cheng, Zhenya Huang, GuanHao Zhao, Yuxiang Guo, Xin Lin, Jinze Wu, Xin Li, Shijin Wang

Automatically generating high-quality mathematical problems that align with educational objectives is a crucial task in NLP-based educational technology. Traditional generation methods focus primarily on textual quality, but they often overlook educational objectives. Moreover, these methods address only single-dimensional, simple question generation, failing to meet complex, multifaceted educational requirements. To address these challenges, we constructed and annotated EduMath, a dataset of 16k mathematical questions with multi-dimensional educational objectives. Based on this dataset, we developed EQGEVAL, which incorporates three evaluation dimensions and is designed to assess the ability of models to generate educational questions. Drawing inspiration from teachers' problem design processes, we propose the Educational Question Planning with self-Reflection (EQPR) method for educational mathematical question generation, following a "plan-evaluate-optimize" approach. Specifically, by combining planning algorithm based on Monte Carlo Tree Search with the generative capabilities of Large Language Models, we continuously optimize questions through iterative feedback. This self-optimization mechanism ensures that the generated questions both fit the educational context and strategically achieve specific basic educational objectives. Through extensive experiments based on EQGEVAL, we have demonstrated that EQPR achieves significant improvements in generating questions that meet multi-dimensional educational objectives.

CLAIM: Mitigating Multilingual Object Hallucination in Large Vision-Language Models with Cross-Lingual Attention Intervention

Zekai Ye, Qiming Li, Xiaocheng Feng, Libo Qin, Yichong Huang, Baohang Li, Kui Jiang, Yang Xiang, Zhirui Zhang, Yunfei Lu, Duyu Tang, Dandan Tu, Bing Qin

Large Vision-Language Models (LVLMs) have demonstrated impressive multimodal abilities but remain prone to multilingual object hallucination, with a higher likelihood of generating responses inconsistent with the visual input when utilizing queries in non-English languages compared to English. Most existing approaches to address these rely on pretraining or fine-tuning, which are resource-intensive. In this paper, inspired by observing the disparities in cross-modal attention patterns across languages, we propose Cross-Lingual Attention Intervention for Mitigating multilingual object hallucination (CLAIM) in LVLMs, a novel near training-free method by aligning attention patterns. CLAIM first identifies language-specific cross-modal attention heads, then estimates language shift vectors from English to the target language, and finally intervenes in the attention outputs during inference to facilitate cross-lingual visual perception capability alignment. Extensive experiments demonstrate that CLAIM achieves an average improvement of 13.56% (up to 30% in Spanish) on the POPE and 21.75% on the hallucination subsets of the MME benchmark across various languages. Further analysis reveals that multilingual attention divergence is most prominent in intermediate layers, highlighting their critical role in multilingual scenarios.

Can LLM Watermarks Robustly Prevent Unauthorized Knowledge Distillation?

Leyi Pan, Aiwei Liu, Shiyu Huang, Yijian LU, Xuming Hu, Lijie Wen, Irwin King, Philip S. Yu

The radioactive nature of Large Language Model (LLM) watermarking enables the detection of watermarks inherited by student models when trained on the outputs of watermarked teacher models, making it a promising tool for preventing unauthorized knowledge distillation. However, the robustness of watermark radioactivity against adversarial actors remains largely unexplored. In this paper, we investigate whether student models can acquire the capabilities of teacher models through knowledge distillation while avoiding watermark inheritance. We propose two categories of watermark removal approaches: pre-distillation removal through untargeted and targeted training data paraphrasing (UP and TP), and post-distillation removal through inference-time watermark neutralization (WN). Extensive experiments across multiple model pairs, watermarking schemes and hyper-parameter settings demonstrate that both TP and WN thoroughly eliminate inherited watermarks, with WN achieving this while maintaining knowledge transfer efficiency and low computational overhead. Given the ongoing deployment of watermarking techniques in production LLMs, these findings emphasize the urgent need for more robust defense strategies.

SkillAggregation: Reference-free LLM-Dependent Aggregation

Guangzhi Sun, Anmol Kagrecha, Potsawee Manakul, Phil Woodland, Mark Gales

Large Language Models (LLMs) are increasingly used to assess NLP tasks due to their ability to generate human-like judgments. Single LLMs were used initially, however, recent work suggests using multiple LLMs as judges yields improved performance. An important step in exploiting multiple judgements is the combination stage, aggregation. Existing methods in NLP either assign equal weight to all LLM judgments or are designed for specific tasks such as hallucination detection. This work focuses on aggregating predictions from multiple systems where no reference labels are available. A new method called SkillAggregation is proposed, which learns to combine estimates from LLM judges without needing additional data or ground truth. It extends the Crowdlayer aggregation method, developed for image classification, to exploit the judge estimates during inference. The approach is compared to a range of standard aggregation methods on HaluEval-Dialogue, TruthfulQA and Chatbot Arena tasks. SkillAggregation outperforms Crowdlayer on all tasks, and yields the best performance over all approaches on the majority of tasks.

M2RC-EVAL: Massively Multilingual Repository-level Code Completion Evaluation

Jiaheng Liu, Ken Deng, Congnan Liu, Jian Yang, Shukai Liu, He Zhu, Peng Zhao, Linzheng Chai, Yanan Wu, Ge Zhang, Yingshui Tan, Zekun Moore Wang, JinKe, Zhaoxiang Zhang, Bangyu Xiang, Guoan Zhang, Wenbo Su, Bo Zheng

Repository-level code completion has drawn great attention in software engineering, and several benchmarks have been introduced. However, existing repository-level code completion benchmarks usually focus on a limited number of languages (<5), which cannot evaluate the general code intelligence abilities across different languages for existing code Large Language Models (LLMs). Besides, the existing benchmarks usually report overall average scores of different languages, where the fine-grained abilities in different completion scenarios are ignored. Therefore, to facilitate the research of code LLMs in multilingual scenarios, we propose a massively multilingual repository-level code completion benchmark covering 18 programming languages (called M2RC-EVAL), and two types of fine-grained annotations (i.e., bucket-level and semantic-level) on different completion scenarios are provided, where we obtain these annotations based on the parsed abstract syntax tree. Moreover, we also curate a massively multilingual instruction corpora M2RC-INSTRUCT dataset to improve the repository-level code completion abilities of existing code LLMs. Comprehensive experimental results demonstrate the effectiveness of our M2RC-EVAL and M2RC-INSTRUCT.

Masks Can be Learned As An Alternative of Experts

Peiyu Liu, Tianwen Wei, Bo Zhu, Xin Zhao, Shuicheng YAN

In this work, we investigate how to sparsify a pre-trained dense large language model into a mixture-of-experts (MoE) architecture for faster inference. Our approach applies mask matrix to the activations for each expert, constrained by L_0 regularization to minimize the number of activated parameters. Starting with all parameters active, the model is progressively sparsified during training, ensuring minimal performance loss. This approach proves more efficient than one-shot sparsification techniques, which typically require significant resources for performance recovery. Moreover, our approach automatically identifies shared, token-specific, and inactive experts, allowing for more efficient allocation of computational resources. Through extensive experiments, we achieve up to 97% performance retention on downstream tasks with only 50% of the feed-forward parameters activated in dense models. Beyond enhancing inference efficiency, this strategy of sharing computational units among experts presents a valuable framework for designing more generalized and efficient MoE architectures, opening avenues for future advancements in expert-based models.

Velocitune: A Velocity-based Dynamic Domain Reweighting Method for Continual Pre-training

Zheheng Luo, Xin Zhang, Xiao Liu, Haoling Li, Yeyun Gong, Qi Chen, Peng CHENG

It is well-known that a diverse corpus is critical for training large language models, which are typically constructed from a mixture of various domains. In general, previous efforts resort to either sampling training data from different domains with static proportions or dynamically adjusting these proportions during training to optimise pretraining performance. However, few methods addressed the complexities of domain-adaptive continual pre-training. To fill this gap, we propose Velocitune, a novel framework that dynamically assesses learning velocity and adjusts data proportions accordingly, favouring slower learning domains while de-emphasising faster learning ones, which is guided by a scaling law to estimate the desired learning goal for each domain with a less associated cost. To evaluate the effectiveness of Velocitune, we conduct experiments on a dataset focused on reasoning tasks with CodeLlama, as well as on a corpus of system commands using Llama3 and Mistral. Velocitune achieves performance gains in both math and code reasoning tasks and command-line generation benchmarks. Further analysis reveals that key factors driving Velocitune's effectiveness include target estimation and data ordering.

FRACTAL: Fine-Grained Scoring from Aggregate Text Labels

Yukti Makhija, Priyanka Agrawal, Rishi Saket, Aravindan Raghuveer

Fine-Tuning of LLMs using RLHF / RLAIF has been shown as a critical step to improve the performance of LLMs in complex generation tasks. These methods typically use response-level human or model feedback for alignment. Recent works indicate that finer sentence or span-level labels provide more accurate and interpretable feedback for LLM optimization. In this work, we propose FRACTAL, a suite of models to disaggregate response-level labels into sentence-level (pseudo-)labels through Multiple Instance Learning (MIL) and Learning from Label Proportions (LLP) formulations, novel usage of prior information, and maximum likelihood calibration. We perform close to 2000 experiments across 6 datasets and 4 tasks that show that FRACTAL can reach up to 93% of the performance of the fully supervised baseline while requiring only around 10% of the gold labels. Furthermore, in a downstream eval, employing step-level pseudo scores in RLHF for a math reasoning task leads to 5% absolute improvement in performance. Our work is the first to develop response-level feedback to sentence-level scoring techniques leveraging sentence-level prior information, along with comprehensive evaluations on multiple tasks as well as end-to-end finetuning evaluations.

KERL: Knowledge-Enhanced Personalized Recipe Recommendation using Large Language Models

Fnu Mohbat, Mohammed J Zaki

Recent advances in large language models (LLMs) and the abundance of food data have resulted in studies to improve food understanding using LLMs. Despite several recommendation systems utilizing LLMs and Knowledge Graphs (KGs), there has been limited research on integrating food related KGs with LLMs. We introduce KERL, a unified system that leverages food KGs and LLMs to provide personalized food recommendations and generates recipes with associated micro-nutritional information. Given a natural language question, KERL extracts entities, retrieves subgraphs from the KG, which are then fed into the LLM as context to select the recipes that satisfy the constraints. Next, our system generates the cooking steps and nutritional information for each recipe. To evaluate our approach, we also develop a benchmark dataset by curating recipe related questions, combined with constraints and personal preferences. Through extensive experiments, we show that our proposed KG-augmented LLM significantly outperforms existing approaches, offering a complete and coherent solution for food recommendation, recipe generation, and nutritional analysis. Our code and benchmark datasets are publicly available at https://github.com/mohbattharani/KERL.

Synergistic Weak-Strong Collaboration by Aligning Preferences

Yizhu Jiao, Xuchao Zhang, Zhaoyang Wang, Yubo Ma, Zhun Deng, Rujia Wang, Chetan Bansal, Saravan Rajmohan, Jiawei Han, Huaxiu Yao

Current Large Language Models excel in general reasoning yet struggle with specialized tasks requiring proprietary or domain-specific knowledge. Fine-tuning large models for every niche application is often infeasible due to black-box constraints and high computational overhead. To address this, we propose a collaborative framework that pairs a specialized weak model with a general strong model. The weak model, tailored to specific domains, produces initial drafts and background information, while the strong model leverages its advanced reasoning to refine these drafts, extending LLMs' capabilities to critical yet specialized tasks. To optimize this collaboration, we introduce a collaborative feedback to fine-tunes the weak model, which quantifies the influence of the weak model's contributions in the collaboration procedure and establishes preference pairs to guide preference tuning of the weak model. We validate our framework through experiments on three domains. We find that the collaboration significantly outperforms each model alone by leveraging complementary strengths. Moreover, aligning the weak model with the collaborative preference further enhances overall performance.

Improving Low-Resource Morphological Inflection via Self-Supervised Objectives

Adam Wiemerslage, Katharina von der Wense

Self-supervised objectives have driven major advances in NLP by leveraging large-scale unlabeled data, but such resources are scarce for many of the world's languages. Surprisingly, they have not been explored much for character-level tasks, where smaller amounts of data have the potential to be beneficial. We investigate the effectiveness of self-supervised auxiliary tasks for morphological inflection -- a character-level task highly relevant for language documentation -- in extremely low-resource settings, training encoder-decoder transformers for 19 languages and 13 auxiliary objectives. Autoencoding yields the best performance when unlabeled data is very limited, while character masked language modeling (CMLM) becomes more effective as data availability increases. Though objectives with stronger inductive biases influence model predictions intuitively, they rarely outperform standard CMLM. However, sampling masks based on known morpheme boundaries consistently improves performance, highlighting a promising direction for low-resource morphological modeling.

Diffusion Directed Acyclic Transformer for Non-Autoregressive Machine Translation

Quan Nguyen-Tri, Cong Dao Tran, Hoang Thanh-Tung

Non-autoregressive transformers (NATs) predict entire sequences in parallel to reduce decoding latency, but they often encounter performance challenges due to the multi-modality problem. A recent advancement, the Directed Acyclic Transformer (DAT), addresses this issue by capturing multiple translation modalities to paths in a Directed Acyclic Graph (DAG). However, the collaboration with the latent variable introduced through the Glancing training (GLAT) is crucial for DAT to attain state-of-the-art performance. In this paper, we introduce Diffusion Directed Acyclic Transformer (Diff-DAT), which serves as an alternative to GLAT as a latent variable introduction for DAT. Diff-DAT offers two significant benefits over the previous approach. Firstly, it establishes a stronger alignment between training and inference. Secondly, it facilitates a more flexible tradeoff between quality and latency.

Unveiling Privacy Risks in LLM Agent Memory

Bo Wang, Weiyi He, Shenglai Zeng, Zhen Xiang, Yue Xing, Jiliang Tang, Pengfei He

Large Language Model (LLM) agents have become increasingly prevalent across various real-world applications. They enhance decision-making by storing private user-agent interactions in the memory module for demonstrations, introducing new privacy risks for LLM agents. In this work, we systematically investigate the vulnerability of LLM agents to our proposed Memory EXTRaction Attack (MEXTRA) under a black-box setting. To extract private information from memory, we propose an effective attacking prompt design and an automated prompt generation method based on different levels of knowledge about the LLM agent. Experiments on two representative agents demonstrate the effectiveness of MEXTRA. Moreover, we explore key factors influencing memory leakage from both the agent designer's and the attacker's perspectives. Our findings highlight the urgent need for effective memory safeguards in LLM agent design and deployment.

Do Language Models Have Semantics? On the Five Standard Positions

Anders Søgaard

We identify five positions on whether large language models (LLMs) and chatbots can be said to exhibit semantic understanding. These positions differ in whether they attribute semantics to LLMs and/or chatbots trained on feedback, what kind of semantics they attribute (inferential or referential), and in virtue of what they attribute referential semantics (internal or external causes). This allows for 2^^4=16 logically possible positions, but we have only seen people argue for five of these. Based on a pairwise comparison of these five positions, we conclude that the better theory of semantics in large language models is, in fact, a sixth combination: Both large language models and chatbots have inferential and referential semantics, grounded in both internal and external causes.

Teaching an Old LLM Secure Coding: Localized Preference Optimization on Distilled Preferences

Mohammad Saqib Hasan, Saikat Chakraborty, Santu Karmaker, Niranjan Balasubramanian

LLM generated code often contains security issues. We address two key challenges in improving secure code generation. First, obtaining high quality training data covering a broad set of security issues is critical. To address this, we introduce a method for distilling a preference dataset of insecure and secure code pairs from frontier LLMs, along with a security reasoning that explains the issues and the fix. The key idea here is to make use of security knowledge sources to devise a systematic prompting strategy that ensures broad coverage. Second, aligning models to secure code requires focusing on localized regions of code. Direct preference optimization methods, like SimPO, are not designed to handle these localized differences and turn out to be ineffective. We address this with a new localized preference optimization algorithm that masks the security related tokens in both the winning (secure) and losing (insecure) responses. To prevent loss in code quality, we also add a regularizer. Evaluations show that both training on our dataset, DiSCo, and the new preference optimization algorithm, LPO, yield substantial reductions in code insecurity while also improving overall code quality. Code and dataset are available at https://github.com/StonyBrookNLP/disco-lpo.

LLM Agents Making Agent Tools

Georg Wölflein, Dyke Ferber, Daniel Truhn, Ognjen Arandjelovic, Jakob Nikolas Kather

Tool use has turned large language models (LLMs) into powerful agents that can perform complex multi-step tasks by dynamically utilising external software components. However, these tools must be implemented in advance by human developers, hindering the applicability of LLM agents in domains demanding large numbers of highly specialised tools, like in life sciences and medicine. Motivated by the growing trend of scientific studies accompanied by public code repositories, we propose ToolMaker, an agentic framework that autonomously transforms papers with code into LLM-compatible tools. Given a GitHub URL and short task description, ToolMaker autonomously installs dependencies and generates code to perform the task, using a closed-loop self-correction mechanism for debugging. To evaluate our approach, we introduce a benchmark comprising 15 complex computational tasks spanning various domains with over 100 unit tests to assess correctness and robustness. Our method correctly implements 80% of the tasks, substantially outperforming current state-of-the-art software engineering agents. ToolMaker therefore is a step towards fully autonomous agent-based scientific workflows.

Cultural Bias Matters: A Cross-Cultural Benchmark Dataset and Sentiment-Enriched Model for Understanding Multimodal Metaphors

Senqi Yang, Dongyu Zhang, Jing Ren, Ziqi Xu, Xiuzhen Zhang, Yiliao Song, Hongfei Lin, Feng Xia

Metaphors are pervasive in communication, making them crucial for natural language processing (NLP). Previous research on automatic metaphor processing predominantly relies on training data consisting of English samples, which often reflect Western European or North American biases. This cultural skew can lead to an overestimation of model performance and contributions to NLP progress. However, the impact of cultural bias on metaphor processing, particularly in multimodal contexts, remains largely unexplored. To address this gap, we introduce MultiMM, a Multicultural Multimodal Metaphor dataset designed for cross-cultural studies of metaphor in Chinese and English. MultiMM consists of 8,461 text-image advertisement pairs, each accompanied by fine-grained annotations, providing a deeper understanding of multimodal metaphors beyond a single cultural domain. Additionally, we propose Sentiment-Enriched Metaphor Detection (SEMD), a baseline model that integrates sentiment embeddings to enhance metaphor comprehension across cultural backgrounds. Experimental results validate the effectiveness of SEMD on metaphor detection and sentiment analysis tasks. We hope this work increases awareness of cultural bias in NLP research and contributes to the development of fairer and more inclusive language models.

Is linguistically-motivated data augmentation worth it?

Ray Groshan, Michael Ginn, Alexis Palmer

Data augmentation, a widely-employed technique for addressing data scarcity, involves generating synthetic data examples which are then used to augment available training data. Researchers have seen surprising success from simple methods, such as random perturbations from natural examples, where models seem to benefit even from data with nonsense words, or data that doesn't conform to the rules of the language. A second line of research produces synthetic data that does in fact follow all linguistic constraints; these methods require some linguistic expertise and are generally more challenging to implement. No previous work has done a systematic, empirical comparison of both linguistically-naive and linguistically-motivated data augmentation strategies, leaving uncertainty about whether the additional time and effort of linguistically-motivated data augmentation work in fact yields better downstream performance. In this work, we conduct a careful and comprehensive comparison of augmentation strategies (both linguistically-naive and linguistically-motivated) for two low-resource languages with different morphological properties, Uspanteko and Arapaho. We evaluate the effectiveness of many different strategies and their combinations across two important sequence-to-sequence tasks for low-resource languages: machine translation and interlinear glossing. We find that linguistically-motivated strategies can have benefits over naive approaches, but only when the new examples they produce are not significantly unlike the training data distribution.

MAGNET: Augmenting Generative Decoders with Representation Learning and Infilling Capabilities

Savya Khosla, Aditi Tiwari, Kushal Kafle, Simon Jenni, Handong Zhao, John Collomosse, Jing Shi

While originally designed for unidirectional generative modeling, decoder-only large language models (LLMs) are increasingly being adapted for bidirectional modeling. However, unidirectional and bidirectional models are typically trained separately with distinct objectives (generation and representation learning). This separation overlooks the opportunity for developing a more versatile language model and for these objectives to complement each other. In this work, we propose MAGNET, a method for adapting decoder-only LLMs to generate robust representations and infill missing text spans. MAGNET employs three self-supervised training objectives and introduces an attention mechanism that combines bidirectional and causal attention, enabling unified training across all objectives. Our results demonstrate that LLMs adapted with MAGNET (1) surpass strong text encoders on token-level and sentence-level representation learning tasks, (2) generate contextually appropriate text infills by leveraging past and future contexts, (3) perform open-ended text generation without excessive repetition of words or phrases, and (4) preserve the knowledge and reasoning capability gained by the LLM during pretraining.

Just a Scratch: Enhancing LLM Capabilities for Self-harm Detection through Intent Differentiation and Emoji Interpretation

Soumitra Ghosh, gopendra Vikram singh, Shambhavi, Sabarna Choudhury, Asif Ekbal

Self-harm detection on social media is critical for early intervention and mental health support, yet remains challenging due to the subtle, context-dependent nature of such expressions. Identifying self-harm intent aids suicide prevention by enabling timely responses, but current large language models (LLMs) struggle to interpret implicit cues in casual language and emojis. This work enhances LLMs' comprehension of self-harm by distinguishing intent through nuanced language–emoji interplay. We present the \textit{C}entennial \textit{E}moji \textit{S}ensitivity \textit{M}atrix (\textit{CESM-100})—a curated set of 100 emojis with contextual self-harm interpretations—and the \textit{S}elf-\textit{H}arm \textit{I}dentification a\textit{N}d intent \textit{E}xtraction with \textit{S}upportive emoji sensitivity (\textit{SHINES}) dataset, offering detailed annotations for self-harm labels, casual mentions (CMs), and serious intents (SIs). Our unified framework: a) enriches inputs using CESM-100; b) fine-tunes LLMs for multi-task learning—self-harm detection (primary) and CM/SI span detection (auxiliary); c) generate explainable rationales for self-harm predictions. We evaluate the framework on three state-of-the-art LLMs—Llama 3, Mental-Alpaca, and MentalLlama—across zero-shot, few-shot, and fine-tuned scenarios. By coupling intent differentiation with contextual cues, our approach commendably enhances LLM performance in both detection and explanation tasks, effectively addressing the inherent ambiguity in self-harm signals. The \textit{SHINES} dataset, \textit{CESM-100} and codebase are publicly available at: https://www.iitp.ac.in/%7eai-nlp-ml/resources.html#SHINES

LLMs Caught in the Crossfire: Malware Requests and Jailbreak Challenges

Haoyang Li, Huan Gao, Zhiyuan Zhao, Zhiyu Lin, Junyu Gao, Xuelong Li

The widespread adoption of Large Language Models (LLMs) has heightened concerns about their security, particularly their vulnerability to jailbreak attacks that leverage crafted prompts to generate malicious outputs. While prior research has been conducted on general security capabilities of LLMs, their specific susceptibility to jailbreak attacks in code generation remains largely unexplored. To fill this gap, we propose MalwareBench, a benchmark dataset containing 3,520 jailbreaking prompts for malicious code-generation, designed to evaluate LLM robustness against such threats. MalwareBench is based on 320 manually crafted malicious code generation requirements, covering 11 jailbreak methods and 29 code functionality categories. Experiments show that mainstream LLMs exhibit limited ability to reject malicious code-generation requirements, and the combination of multiple jailbreak methods further reduces the model's security capabilities: specifically, the average rejection rate for malicious content is 60.93%, dropping to 39.92% when combined with jailbreak attack algorithms. Our work highlights that the code security capabilities of LLMs still pose significant challenges.

T-REG: Preference Optimization with Token-Level Reward Regularization

Wenxuan Zhou, Shujian Zhang, Lingxiao Zhao, Tao Meng

Reinforcement Learning from Human Feedback (RLHF) has been pivotal in enabling Large Language Models (LLMs) to effectively follow instructions and produce meaningful alignment by leveraging human preference data. Traditionally, RLHF involves generating responses to a query and using a separate reward model to assign a score to the entire completion. This approach, however, presents challenges, as it provides a single, sparse reward at the end of a sequence, making optimization difficult for the model, in which both training and generation occur auto-regressively at token levels. While recent methods have attempted to address this by assigning token-level discrete or continuous rewards, these often rely on either a trained credit assignment model or AI annotators, which raises concerns about the quality and reliability of the token-level rewards. In this paper, we propose T-REG, which utilizes both sequence-level and token-level rewards for preference optimization. T-REG employs self-generated token-level rewards, derived through opposite prompting, as a weak supervision signal to guide the model in distributing sequence-level rewards at the token level, thereby achieving more effective token-level credit assignment and improving alignment performance. Experiments on the instruction following benchmarks, including Alpaca Eval 2 and Arena-Hard, show that our method consistently outperforms baseline methods by up to 3.8% and 4.4%, respectively.

Persistent Homology of Topic Networks for the Prediction of Reader Curiosity

Manuel D.S. Hopp, Vincent Labatut, Arthur Amalvy, Richard Dufour, Hannah Stone, Hayley K Jach, Kou Murayama

Reader curiosity, the drive to seek information, is crucial for textual engagement, yet remains relatively underexplored in NLP. Building on Loewenstein's Information Gap Theory, we introduce a framework that models reader curiosity by quantifying semantic information gaps within a text's semantic structure. Our approach leverages BERTopic-inspired topic modeling and persistent homology to analyze the evolving topology (connected components, cycles, voids) of a dynamic semantic network derived from text segments, treating these features as proxies for information gaps. To empirically evaluate this pipeline, we collect reader curiosity ratings from participants (*n* = 49) as they read S. Collins's "The Hunger Games" novel. We then use the topological features from our pipeline as independent variables to predict these ratings, and experimentally show that they significantly improve curiosity prediction compared to a baseline model (73% vs. 30% explained deviance), validating our approach. This pipeline offers a new computational method for analyzing text structure and its relation to reader engagement.

Hierarchical-Task-Aware Multi-modal Mixture of Incremental LoRA Experts for Embodied Continual Learning

Ziqi Jia, Anmin Wang, Xiaoyang Qu, Xiaowen Yang, Jianzong Wang

Previous continual learning setups for embodied intelligence focused on executing low-level actions based on human commands, neglecting the ability to learn high-level planning and multi-level knowledge. To address these issues, we propose the Hierarchical Embodied Continual Learning Setups (HEC) that divide the agent's continual learning process into two layers: high-level instructions and low-level actions, and define five embodied continual learning sub-setups. Building on these setups, we introduce the Task-aware \textbf{M}ixture \textbf{o}f \textbf{I}ncremental \textbf{L}oRA \textbf{E}xperts (Task-aware MoILE) method. This approach achieves task recognition by clustering visual-text embeddings and uses both a task-level router and a token-level router to select the appropriate LoRA experts. To effectively address the issue of catastrophic forgetting, we apply Singular Value Decomposition (SVD) to the LoRA parameters obtained from prior tasks, preserving key components while orthogonally training the remaining parts. The experimental results show that our method stands out in reducing the forgetting of old tasks compared to other methods, effectively supporting agents in retaining prior knowledge while continuously learning new tasks.

Length-Induced Embedding Collapse in PLM-based Models

Yuqi Zhou, Sunhao Dai, Zhanshuo Cao, Xiao Zhang, Jun Xu

Text embeddings from PLM-based models enable a wide range of applications, yet their performance often degrades on longer texts. In this paper, we introduce a phenomenon we call \textbf{Length Collapse}, where embeddings of longer texts tend to cluster together. This clustering results in a distributional inconsistency between the embeddings of short and long texts. We further investigate how these differences contribute to the performance decline observed with longer texts across various downstream tasks. Through a rigorous theoretical analysis of the self-attention mechanism, which acts as a low-pass filter in PLM-based models, we demonstrate that as text length increases, the strength of low-pass filtering intensifies, causing embeddings to retain more low-frequency components. As a result, input token features become more similar, leading to clustering and ultimately the collapse of embeddings for longer texts. To address this issue, we propose a simple method, TempScale, which mitigates the Length Collapse phenomenon. By narrowing the gap in low-pass filtering rates between long and short texts, TempScale ensures more consistent embeddings across different text lengths. This approach leads to performance improvements of \textbf{0.94\%} on MTEB and \textbf{1.10\%} on LongEmbed, which focuses specifically on long-context retrieval, providing strong evidence for the validity of our analysis. The source code is available at \textcolor{blue}{}.

Structure-aware Domain Knowledge Injection for Large Language Models

Kai Liu, Ze Chen, Zhihang Fu, Wei Zhang, Rongxin Jiang, Fan Zhou, Yaowu Chen, Yue Wu, Jieping Ye

This paper introduces a pioneering methodology, termed StructTuning, to efficiently transform foundation Large Language Models (LLMs) into domain specialists. It significantly reduces the training corpus needs to a mere 5% while achieving an impressive 100% of traditional knowledge injection performance. Motivated by structured human education, we propose a novel two-stage strategy for knowledge injection and alignment: Structure-aware Continual Pre-Training (SCPT) and Structure-aware Supervised Fine-Tuning (SSFT). In the SCPT phase, we automatically extract the domain knowledge taxonomy and reorganize the training corpora, enabling LLMs to effectively link textual segments to targeted knowledge points within the taxonomy. In the SSFT phase, we explicitly prompt models to elucidate the underlying knowledge structure in their outputs, leveraging the structured domain insight to address practical problems. Our ultimate method was extensively evaluated across model architectures and scales on LongBench and MMedBench datasets, demonstrating superior performance against other knowledge injection methods. We also explored our method's scalability across different training corpus sizes, laying the foundation to enhance domain-specific LLMs with better data utilization.

PolyNarrative: A Multilingual, Multilabel, Multi-domain Dataset for Narrative Extraction from News Articles

Nikolaos Nikolaidis, Nicolas Stefanovitch, Purificação Silvano, Dimitar Iliyanov Dimitrov, Roman Yangarber, Nuno Guimarães, Elisa Sartori, Ion Androutsopoulos, Preslav Nakov, Giovanni Da San Martino, Jakub Piskorski

We present polyNarrative, a new multilingual dataset of news articles, annotated for narratives. Narratives are overt or implicit claims, recurring across articles and languages, promoting a specific interpretation or viewpoint on an ongoing topic, often propagating mis/disinformation. We developed two-level taxonomies with coarse- and fine-grained narrative labels for two domains: (i) climate change and (ii) the military conflict between Ukraine and Russia. We collected news articles in four languages (Bulgarian, English, Portuguese, and Russian) related to the two domains and manually annotated them at the paragraph level. We make the dataset publicly available, along with experimental results of several strong baselines that assign narrative labels to news articles at the paragraph or the document level. We believe that this dataset will foster research in narrative detection and enable new research directions towards more multi-domain and highly granular narrative related tasks.

Re³Syn: A Dependency-Based Data Synthesis Framework for Long-Context Post-training

Zhiyang Zhang, Ziqiang Liu, Huiming Wang, Renke Shan, Li Kuang, Lu Wang

An important trend in the realm of large language models (LLMs) is the development of longer context windows. However, training LLMs with long context windows to acquire the capability of effectively modeling lengthy inputs is often hindered by the scarcity of naturally long-context data. Existing methods for constructing long-context data by concatenating short documents have overlooked a crucial characteristic of long-context data quality, namely semantic dependency. In this paper, we propose a novel framework called Retrieval, Dependency Recognition, and Reorder for data synthesis (Re^{3}Syn), which leverages semantic similarity to retrieve relevant documents and form several batches. Within each batch, the framework comprehensively recognizes dependency and utilizes them, along with a reorder algorithm, to organize the short documents into coherent long-context data. Comprehensive experiment on multiple benchmarks indicate that the data generated by the Re^{3}Syn has longer dependencies and significantly enhances the model's long-context capabilities. For reproducibility, we will release our codebase upon acceptance.

Merge Hijacking: Backdoor Attacks to Model Merging of Large Language Models

Zenghui Yuan, Yangming Xu, Jiawen Shi, Pan Zhou, Lichao Sun

Model merging for Large Language Models (LLMs) directly fuses the parameters of different models finetuned on various tasks, creating a unified model for multi-domain tasks. However, due to potential vulnerabilities in models available on open-source platforms, model merging is susceptible to backdoor attacks. In this paper, we propose \textit{Merge Hijacking}, the first backdoor attack targeting model merging in LLMs. The attacker constructs a malicious upload model and releases it. Once a victim user merges it with any other models, the resulting merged model inherits the backdoor while maintaining utility across tasks. Merge Hijacking defines two main objectives—effectiveness and utility—and achieves them through four steps. Extensive experiments demonstrate the effectiveness of our attack across different models, merging algorithms, and tasks. Additionally, we show that the attack remains effective even when merging real-world models. Moreover, our attack demonstrates robustness against two inference-time defenses (Paraphrasing and CLEANGEN) and one training-time defense (Fine-pruning).

CFBench: A Comprehensive Constraints-Following Benchmark for LLMs

Tao Zhang, ChengLIn Zhu, Yanjun Shen, Wenjing Luo, Yan Zhang, Hao Liang, Tao Zhang, Fan Yang, Mingan Lin, Yujing Qiao, Weipeng Chen, Bin CUI, Wentao Zhang, Zenan Zhou

The adeptness of Large Language Models (LLMs) in comprehending and following natural language instructions is critical for their deployment in sophisticated real-world applications. Existing evaluations mainly focus on fragmented constraints or narrow scenarios, but they overlook the comprehensiveness and authenticity of constraints from the user's perspective. To bridge this gap, we propose CFBench, a large-scale Chinese Comprehensive Constraints Following Benchmark for LLMs, featuring 1,000 curated samples that cover more than 200 real-life scenarios and over 50 NLP tasks. CFBench meticulously compiles constraints from real-world instructions and constructs an innovative systematic framework for constraint types, which includes 10 primary categories and over 25 subcategories, and ensures each constraint is seamlessly integrated within the instructions. To make certain that the evaluation of LLM outputs aligns with user perceptions, we propose an advanced methodology that integrates multi-dimensional assessment criteria with requirement prioritization, covering various perspectives of constraints, instructions, and requirement fulfillment. Evaluating current leading LLMs on CFBench reveals substantial room for improvement in constraints following, and we further investigate influencing factors and enhancement strategies. The data and code will be made available.

ConsistencyChecker: Tree-based Evaluation of LLM Generalization Capabilities

Zhaochen Hong, Haofei Yu, Jiaxuan You

Evaluating Large Language Models (LLMs) requires effective methods to assess semantic consistency across multiple reversible transformations. Traditional self-consistency methods often fail to capture subtle semantic errors in multi-step tasks. We introduce ConsistencyChecker, a tree-based evaluation framework that measures LLMs' ability to preserve semantic consistency during reversible transformation processes, sidestepping benchmark data contamination issues. Our approach constructs self-consistency trees where nodes represent text states after transformations (e.g., translation, code modification, paraphrasing) and edges represent pairs of opposite transformations. By analyzing semantic preservation between nodes at different tree depths, ConsistencyChecker quantifies model reliability without requiring manually annotated reference data. Experiments demonstrate that ConsistencyChecker reliably measures generalization abilities across models from 1.5B to 72B parameters. On translation tasks, GPT-4o Mini achieves the highest L3 consistency score of 98.0%. For code generation, Qwen 2.5 32B leads with 85.1% semantic consistency at L3. Results show Pearson correlation greater than 0.7 between our embedding-based scores and WMT 2024 rankings on 4 out of 5 shared language pairs, validating the method's effectiveness for benchmarking LLM performance without constructing new datasets.

Diverse AI Feedback For Large Language Model Alignment

Tianshu Yu

In recent research on large language models (LLMs), there has been a growing emphasis on aligning these models with human values to reduce the impact of harmful content. However, current alignment methods often rely solely on singular forms of feedback, such as preferences, annotated labels, or natural language critiques, overlooking the potential advantages of combining these feedback types. This limitation leads to suboptimal performance, even when ample training data is available. In this paper, we introduce Diverse AI Feedback (DAIF) as a novel method to enhance LLM alignment with diverse feedback mechanisms. Our approach involves collecting three distinct types of feedback tailored to problems of varying difficulty levels within the training dataset. Specifically, we exploit critique feedback for easy problems, refinement feedback for medium problems, and preference feedback for hard problems. By training our model with this diversified feedback, we achieve enhanced alignment performance while using less training data. To assess the effectiveness of DAIF, we evaluate it against previous methods in three downstream tasks: question answering, dialog generation, and text summarization. Experimental results demonstrate that DAIF achieves superior performance even with a smaller training dataset.

Session5: V-Gather Find (Monday 28 July, 18:00-19:30,)

Emo Pillars: Knowledge Distillation to Support Fine-Grained Context-Aware and Context-Less Emotion Classification

Alexander Shvets

Most datasets for sentiment analysis lack context in which an opinion was expressed, often crucial for emotion understanding, and are mainly limited by a few emotion categories. Foundation large language models (LLMs) like GPT-4 suffer from over-predicting emotions and are too resource-intensive. We design an LLM-based data synthesis pipeline and leverage a large model, Mistral-7b, for the generation of training examples for more accessible, lightweight BERT-type encoder models. We focus on enlarging the semantic diversity of examples and propose grounding the generation into a corpus of narratives to produce non-repetitive story-character-centered utterances with unique contexts over 28 emotion classes. By running 700K inferences in 450 GPU hours, we contribute with the dataset of 100K contextual and also 300K context-less examples to cover both scenarios. We use it for fine-tuning pre-trained encoders, which results in several Emo Pillars models. We show that Emo Pillars models are highly adaptive to new domains when tuned to specific tasks such as GoEmotions, ISEAR, IEMOCAP, and EmoContext, reaching the SOTA performance on the first three. We also validate our dataset, conducting statistical analysis and human evaluation, and confirm the success of our measures in utterance diversification (although less for the neutral class) and context personalization, while pointing out the need for improved handling of out-of-taxonomy labels within the pipeline.

Translate With Care: Addressing Gender Bias, Neutrality, and Reasoning in Large Language Model Translations

Pardis Sadat Zahraei, Ali Emami

Addressing gender bias and maintaining logical coherence in machine translation remains challenging, particularly when translating between natural gender languages, like English, and genderless languages, such as Persian, Indonesian, and Finnish. We introduce the Translate-with-Care (TWC) dataset, comprising 3,950 challenging scenarios across six low- to mid-resource languages, to assess translation systems' performance. Our analysis of diverse technologies, including GPT-4, mBART-50, NLLB-200, and Google Translate, reveals a universal struggle in translating genderless content, resulting in gender stereotyping and reasoning errors. All models preferred masculine pronouns when gender stereotypes could influence choices. Google Translate and GPT-4 showed particularly strong bias, favoring male pronouns 4-6 times more than feminine ones in leadership and professional success contexts. Fine-tuning mBART-50 on TWC substantially resolved these biases and errors, led to strong generalization, and surpassed proprietary LLMs while remaining open-source. This work emphasizes the need for targeted approaches to gender and semantic coherence in machine translation, particularly for genderless languages, contributing to more equitable and accurate translation systems.

Talking Point based Ideological Discourse Analysis in News Events

Nishanth Sridhar Nakshatri, Nikhil Mehta, Siyi Liu, Sihao Chen, Daniel Hopkins, Dan Roth, Dan Goldwasser

Analyzing ideological discourse even in the age of LLMs remains a challenge, as these models often struggle to capture the key elements that shape real-world narratives. Specifically, LLMs fail to focus on characteristic elements driving dominant discourses and lack the ability to integrate contextual information required for understanding abstract ideological views. To address these limitations, we propose a framework motivated by the theory of ideological discourse analysis that constructs a vocabulary of repeating "talking points" from a corpus of news articles related to real-world events. Our framework contextualizes these talking points by analyzing the interaction between entities, their roles, and media frames to generate ideology-specific viewpoints (or partisan perspectives). We evaluate our framework's ability to generate these perspectives through automated tasks— ideology and partisan classification tasks, supplemented by human validation. Additionally, we demonstrate straightforward applicability of our framework in creating event snapshots, a visual way of interpreting event discourse. We release resulting dataset and model to the community to support further research.

ConceptEdit: Conceptualization-Augmented Knowledge Editing in Large Language Models for Commonsense Reasoning

Liyu Zhang, Weiqi Wang, Tianqing Fang, Yangqiu Song

Knowledge Editing (KE) aims to adjust a Large Language Model's (LLM) internal representations and parameters to correct inaccuracies and improve output consistency without incurring the computational expense of re-training the entire model. However, editing commonsense knowledge still faces difficulties, including limited knowledge coverage in existing resources, the infeasibility of annotating labels for an overabundance of commonsense knowledge, and the strict knowledge formats of current editing methods. In this paper, we address these challenges by presenting ConceptEdit, a framework that integrates conceptualization and instantiation into the KE pipeline for LLMs to enhance their commonsense reasoning capabilities. ConceptEdit dynamically diagnoses implausible commonsense knowledge within an LLM using another verifier LLM and augments the source knowledge to be edited with conceptualization for stronger generalizability. Experimental results demonstrate that LLMs enhanced with ConceptEdit successfully generate commonsense knowledge with improved plausibility compared to other baselines and achieve stronger performance across multiple question answering benchmarks.

SHARP: Unlocking Interactive Hallucination via Stance Transfer in Role-Playing LLMs

Chuyi Kong, Ziyang Luo, Hongzhan Lin, Zhiyuan Fan, Yaxin Fan, Yuxi SUN, Jing Ma

The advanced role-playing capabilities of Large Language Models (LLMs) have enabled rich interactive scenarios, yet existing research in social interactions neglects hallucination while struggling with poor generalizability and implicit character fidelity judgments. To bridge this gap, motivated by human behaviour, we introduce a generalizable and explicit paradigm for uncovering interactive patterns of LLMs across diverse worldviews. Specifically, we first define interactive hallucination through stance transfer, then construct SHARP, a benchmark built by extracting relations from commonsense knowledge graphs and utilizing LLMs' inherent hallucination properties to simulate multi-role interactions. Extensive experiments confirm our paradigm's effectiveness and stability, examine the factors that influence these metrics, and challenge conventional hallucination mitigation solutions. More broadly, our work reveals a fundamental limitation in popular post-training methods for role-playing LLMs: the tendency to obscure knowledge beneath style, resulting in monotonous yet human-like behaviors—interactive hallucination.

GlyphPattern: An Abstract Pattern Recognition for Vision-Language Models

Zixuan Wu, Yoolim Kim, Carolyn Jane Anderson

Vision-Language Models (VLMs) have made rapid progress in reasoning across visual and textual data. While VLMs perform well on vision tasks that they are trained on, our results highlight key challenges in abstract pattern recognition. We present GlyphPattern, a 954 item dataset that pairs 318 human-written descriptions of visual patterns from 40 writing systems with three visual presentation styles. GlyphPattern evaluates abstract pattern recognition in VLMs, requiring models to understand and judge natural language descriptions of visual patterns. GlyphPattern patterns are drawn from a large-scale cognitive science investigation of human writing systems; as a result, they are rich in spatial reference and compositionality. Our experiments show that GlyphPattern is challenging for state-of-the-art VLMs (GPT-4o achieves only 55% accuracy), with marginal gains from few-shot prompting. Our detailed analysis reveals errors at multiple levels, including visual processing, natural language understanding, and pattern generalization.

LogicGame: Benchmarking Rule-Based Reasoning Abilities of Large Language Models

Jiayi Gui, Yiming Liu, Jiale Cheng, Xiaotao Gu, Xiao Liu, Hongning Wang, Yuxiao Dong, Jie Tang, Minlie Huang

Large Language Models (LLMs) have demonstrated notable capabilities across various tasks, showcasing complex problem-solving abilities. Understanding and executing complex rules, along with multi-step planning, are fundamental to logical reasoning and critical for practical LLM agents and decision-making systems. However, evaluating LLMs as effective rule-based executors and planners remains underexplored. In this paper, we introduce LogicGame, a novel benchmark designed to evaluate the comprehensive rule understanding, execution, and planning capabilities of LLMs. Unlike traditional benchmarks, LogicGame provides diverse games that contain a series of rules with an initial state, requiring models to comprehend and apply predefined regulations to solve problems. We create simulated scenarios in which models execute or plan operations to achieve specific outcomes. These game scenarios are specifically designed to distinguish logical reasoning from mere knowledge by relying exclusively on predefined rules. This separation allows for a pure assessment of rule-based reasoning capabilities. The evaluation considers not only final outcomes but also intermediate steps, providing a comprehensive assessment of model performance. Moreover, these intermediate steps are deterministic and can be automatically verified. LogicGame defines game scenarios with varying difficulty levels, from simple rule applications to complex reasoning chains, in order to offer a precise evaluation of model performance on rule understanding and multi-step execution. Utilizing LogicGame, we test various LLMs and identify notable shortcomings in their rule-based logical reasoning abilities.

Towards Efficient LLM Grounding for Embodied Multi-Agent Collaboration

Yang Zhang, Shixin Yang, Chenjia Bai, Fei Wu, Xiu Li, Zhen Wang, Xuelong Li

Grounding the reasoning ability of large language models (LLMs) for embodied tasks is challenging due to the complexity of the physical world. Especially, LLM planning for multi-agent collaboration requires communication of agents or credit assignment as the feedback to re-adjust the proposed plans and achieve effective coordination. However, existing methods that overly rely on physical verification or self-reflection suffer from excessive and inefficient querying of LLMs. In this paper, we propose a novel framework for multi-agent collaboration that introduces Reinforced Advantage feedback (ReAd) for efficient self-refinement of plans. Specifically, we perform critic regression to learn a sequential advantage function from LLM-planned data, and then treat the LLM planner as an optimizer to generate actions that maximize the advantage function. It endows the LLM with the foresight to discern whether the action contributes to accomplishing the final task. We provide theoretical analysis by extending advantage-weighted regression in reinforcement learning to multi-agent systems. Experiments on Overcooked-AI and a difficult variant of RoCoBench show that ReAd surpasses baselines in success rate, and also significantly decreases the interaction steps of agents and query rounds of LLMs, demonstrating its high efficiency for grounding LLMs. More results are given at https://read-llm.github.io/.

Critic-CoT: Boosting the Reasoning Abilities of Large Language Model via Chain-of-Thought Critic

Xin Zheng, Jie Lou, Boxi Cao, Xueru Wen, Yuqiu Ji, Hongyu Lin, Yaojie Lu, Xianpei Han, Debing Zhang, Le Sun

Self-critic has become a crucial mechanism for enhancing the reasoning performance of LLMs. However, current approaches mainly involve basic prompts for intuitive instance-level feedback, which resembles System-1 processes and limits the reasoning capabilities. Moreover, there is a lack of in-depth investigations into the relationship between LLM's ability to criticize and its task-solving performance. To address these issues, we propose Critic-CoT, a novel framework that pushes LLMs toward System-2-like critic capability. Through a step-wise CoT reasoning paradigm and the automatic construction of weak-supervision data without human annotation, Critic-CoT enables LLMs to engage in slow, analytic self-critique and refinement, thereby improving their reasoning abilities. Experiments on GSM8K and MATH and out-of-domain evaluation demonstrate that our enhanced model significantly boosts task-solving performance by filtering out invalid solutions or iterative refinement. Furthermore, we investigate the intrinsic correlation between critique and task-solving abilities within LLMs, discovering that these abilities can mutually reinforce each other rather than conflict.

Measuring Bias and Agreement in Large Language Model Presupposition Judgments

Katherine Atwell

Identifying linguistic bias in text requires understanding what is said and what is meant. This requires going beyond what is being asserted directly, and determining what is presupposed. Large language models (LLMs) represent a potential automatic approach for identifying presupposed content, but it is unknown how well LLM judgments correspond to human judgments. Further, LLMs may exhibit their own biases in determining what is presupposed. To study this empirically, we prompt multiple LLMs to make presupposition judgments for texts of varying domains from three different human-labeled datasets. We calculate the agreement between LLMs and human raters, and find several linguistic factors associated with fluctuations in human-model agreement. We also observe discrepancies in agreement scores that indicate potential biases from LLMs, related to gender and political ideology.

Adaptive-VP: A Framework for LLM-Based Virtual Patients that Adapts to Trainees' Dialogue to Facilitate Nurse Communication Training

Keyeun Lee, Seolhee Lee, Esther Hehsun Kim, Yena Ko, Jinsu Eun, Dahee Kim, Hyewon Cho, Haiyi Zhu, Robert E. Kraut, Eunyoung E. Suh, Eun-mee Kim, Hajin Lim

Effective nurse communication is vital for high-quality patient care, yet standardized patient (SP) simulations are costly and inflexible, and virtual patient (VP) systems often fail to adapt dynamically to trainees' communication skills. When a trainee responds ineffectively, the VP should escalate hostility or become uncooperative, yet most systems lack this adaptive capability. To address this, we introduce Adaptive-VP, an LLM-based framework that flexibly adjusts VP dialogue. Our framework includes a pipeline for constructing clinically grounded yet adaptable VP cases and implements a modular structure for evaluating trainee communication and dynamically adjusting VP responses, while ensuring learner safety. Specifically, we applied and validated Adaptive-VP by simulating challenging patient conversations. Automated evaluation of the corpus from practicing nurses showed that our evaluation mechanism mirrored real-world communication efficiency levels. Human evaluation with expert nurses further confirmed that Adaptive-VP generates more realistic and natural interactions than existing approaches, demonstrating its potential as a scalable and effective tool for nursing communication.

LTRAG: Enhancing autoformalization and self-refinement for logical reasoning with Thought-Guided RAG

Ruikang Hu, Shaoyu Lin, Yeliang Xiu, Yongmei Liu

Logical reasoning is fundamental to intelligent systems. Large language models (LLMs) have demonstrated promise in natural language (NL) reasoning, especially with techniques like chain-of-thought (CoT) prompting. Neuro-symbolic methods like Logic-LM and LINC further enhance performance on challenging datasets FOLIO and AR-LSAT by integrating formalization with LLMs and symbolic solvers, and possibly refinement with LLMs. However, these methods still struggle with the accurate formalization of complex NL problems. In this paper, we introduce LTRAG, a framework to enhance autoformalization and self-refinement for logical reasoning with Retrieval-Augmented Generation (RAG), by building knowledge bases of thought-guided examples. Experimental results on FOLIO and AR-LSAT show that LTRAG consistently outperforms Logic-LM and LINC across different models. On GPT-4 and AR-LSAT, it achieves an accuracy gain of 13% over Logic-LM.

Do Language Models Understand the Cognitive Tasks Given to Them? Investigations with the N-Back Paradigm

Xiaoyang Hu, Richard Lewis

Cognitive tasks originally developed for humans are now increasingly used to study language models. While applying these tasks is often straightforward, interpreting their results can be challenging. In particular, when a model underperforms, it is often unclear whether this results from a limitation in the cognitive ability being tested or a failure to understand the task itself. A recent study argues that GPT 3.5's declining performance on 2-back and 3-back tasks reflects a working memory capacity limit similar to humans (Gong et al., 2024). By analyzing a range of open-source language models of varying performance levels on these tasks, we show that the poor performance is due at least in part to a limitation in task comprehension and task set maintenance. We challenge the best-performing model with progressively harder versions of the task (up to 10-back) and experiment with alternative prompting strategies, before analyzing model attentions. Our larger aim is to contribute to the ongoing conversation around refining methodologies for the cognitive evaluation of language models.

Making RALM Robust to Irrelevant Contexts via Layer Knowledge Guided Attention

Weijie Shi, Hao Chen, Jiaming Li, Yao Zhao, Yazhong Zhang, Qijin Chen, Jipeng Zhang, Ruiyuan Zhang, Jia Zhu, Jiajie Xu, Xiaofang Zhou

Retrieval-augmented language models (RALMs) aim to incorporate external knowledge to address the issues of factual hallucination and knowledge obsolescence faced by large language models (LLMs). Inevitably, the retrieved passages based on similarity search may be irrelevant to the given question, and the aggregation of these passages can confuse the model to give a correct answer. To improve the performance of RALM in such conditions, we propose layer-knowledge guided attention for RALMs, which harnesses the layer-wise knowledge of LLMs to optimize per-layer attention on useful passages, making the model pay attention to the most relevant content and ignore irrelevant ones. Specifically, we first systematically study LLM's attention patterns and their relationship with the accuracy of RALM responses, where middle-focus attentions play a crucial role in selectively gathering relevant information. Based on this, a layer-wise passage estimator leverages the varied knowledge encoded across LLM layers to assess not only passage relevance scores but also associated confidences. Finally, a relevance-aware passage fusion enables selective attention to relevant passages, mitigating distractibility and positional bias of causal attention. Experiments show that our method outperforms existing methods on RALM benchmarks.

Large Language Models as Sign Language Interfaces: Mitigating the Requests of Deaf Users of LLMs in a Hearing-Centric World

Mert Inan, Anthony Sicilia, Malihe Alikhani

Deaf or Hard-of-Hearing (DHH) individuals use Large Language Models (LLMs) in unique ways and request to incorporate sign language grammar and Deaf culture during the training of these models, in addition to video-based sign language input capabilities. Yet, developers of instruct-tuned LLMs have not paid attention to these requests. Instead, special translation models are developed for sign languages (SLs), diminishing the needs of signers to a simple lack of communication between hearing and Deaf communities. In this paper, we take an orthogonal approach to these traditional methods of studying SLs. To meet the requests of Deaf users of LLMs, we look at the sign language processing (SLP) algorithm from a theoretical lens, then introduce the first text-based and multimodal LLMs. We propose new prompting and fine-tuning strategies for text-based and multimodal SLP, incorporating sign linguistic rules and conventions. We test the generalization of these models to other SLP tasks, showing LLMs can process signs while still being adept at spoken language tasks. Our code and model checkpoints will be open-source. We will update our model suite as newer open-source LLMs, datasets, and SLP tasks become available.

Unveiling Confirmation Bias in Chain-of-Thought Reasoning

Yue Wan, Xiaowei Jia, Xiang Lorraine Li

Chain-of-thought (CoT) prompting has been widely adopted to enhance the reasoning ca- pabilities of large language models (LLMs). However, the effectiveness of CoT reasoning is inconsistent across tasks with different reason- ing types. This work presents a novel perspec- tive to understand CoT behavior through the lens of confirmation bias in cognitive psychol- ogy. Specifically, we examine how model in- ternal beliefs, approximated by direct question- answering probabilities, affect both reasoning generation (Q -> R) and reasoning-guided an- swer prediction (QR -> A) in CoT. By decom- posing CoT into a two-stage process, we con- duct a thorough correlation analysis in model beliefs, rationale attributes, and stage-wise per- formance. Our results provide strong evidence of confirmation bias in LLMs, such that model beliefs not only skew the reasoning process but also influence how rationales are utilized for answer prediction. Furthermore, the interplay between task vulnerability to confirmation bias and the strength of beliefs also provides expla- nations for CoT effectiveness across reasoning tasks and models. Overall, this study provides a valuable insight for the needs of better prompt- ing strategies that mitigate confirmation bias to enhance reasoning performance. Code is avail- able at https://github.com/yuewan2/biasedcot.

Taxonomizing Representational Harms using Speech Act Theory

Hannah Washington, Emily Corvi, Stefanie Reed, Chad Atalla, Alexandra Chouldechova, P. Alex Dow, Jean Garcia-Gathright, Nicholas J Pangakis, Emily Sheng, Dan Vann, Matthew Vogel, Hanna Wallach

Representational harms are widely recognized among fairness-related harms caused by generative language systems. However, their definitions are commonly under-specified. We present a framework, grounded in speech act theory (Austin, 1962), that conceptualizes representational harms caused by generative language systems as the perlocutionary effects (i.e., real-world impacts) of particular types of illocutionary acts (i.e., system behaviors). Building on this argument and drawing on relevant literature from linguistic anthropology and sociolinguistics, we provide new definitions for stereotyping, demeaning, and erasure. We then use our framework to develop a granular taxonomy of illocutionary acts that cause representational harms, going beyond the high-level taxonomies presented in previous work. We also discuss the ways that our framework and taxonomy can support the development of valid measurement instruments. Finally, we demonstrate the utility of our framework and taxonomy via a case study that engages with recent conceptual debates about what constitutes a representational harm and how such harms should be measured.

Scaling Laws for Multilingual Language Models

Yifei He, Alon Benhaim, Barun Patra, Praneetha Vaddamanu, Sanchit Ahuja, Parul Chopra, Vishrav Chaudhary, Han Zhao, Xia Song

We propose a novel scaling law for general-purpose decoder-only language models (LMs) trained on multilingual data, tackling the problem of balancing languages during multilingual pretraining. A primary challenge in studying multilingual scaling is the difficulty of analyzing individual language performance due to cross-lingual transfer. To tackle this, we shift the focus from individual languages to language families. We introduce and validate a hypothesis that the test cross-entropy loss for each language family is determined solely by its own sampling ratio, independent of other languages in the mixture. This insight simplifies the complexity of multilingual scaling and make the analysis scalable to an arbitrary number of languages. Building on this hypothesis, we derive a power-law relationship that links performance with dataset size, model size and sampling ratios. This relationship enables us to predict performance across various combinations of the above three quantities, and derive the optimal sampling ratios at different model scales. To demonstrate the effectiveness and accuracy of our proposed scaling law, we perform a large-scale empirical study, training more than 100 models on 23 languages spanning 5 language families. Our experiments show that the optimal sampling ratios derived from small models (85M parameters) generalize effectively to models that are several orders of magnitude larger (1.2B parameters), offering a resource-efficient approach for multilingual LM training at scale.

Vision Language Model Helps Private Information De-Identification in Vision Data

Tiejin Chen, Pingzhi Li, Kaixiong Zhou, Tianlong Chen, Hua Wei

Visual Language Models (VLMs) have gained significant popularity due to their remarkable ability. While various methods exist to enhance privacy in text-based applications, privacy risks associated with visual inputs remain largely overlooked such as Protected Health Information (PHI) in medical images. To tackle this problem, two key tasks: accurately localizing sensitive text and processing it to ensure privacy protection should be performed. To address this issue, we introduce VisShield (Vision Privacy Shield), an end-to-end framework designed to enhance the privacy awareness of VLMs. Our framework consists of two key components: a specialized instruction-tuning dataset OPTIC (Optical Privacy Text Instruction Collection) and a tailored training methodology. The dataset provides diverse privacy-oriented prompts that guide VLMs to perform targeted Optical Character Recognition (OCR) for precise localization of sensitive text, while the training strategy ensures effective adaptation of VLMs to privacy-preserving tasks. Specifically, our approach ensures that VLMs recognize privacy-sensitive text and output precise bounding boxes for detected entities, allowing for effective masking of sensitive information. Extensive experiments demonstrate that our framework significantly outperforms existing approaches in handling private information, paving the way for privacy-preserving applications in vision-language models.

Unveiling Privacy Risks in Multi-modal Large Language Models: Task-specific Vulnerabilities and Mitigation Challenges

Tiejin Chen, Pingzhi Li, Kaixiong Zhou, Tianlong Chen, Hua Wei

Privacy risks in text-only Large Language Models (LLMs) are well studied, particularly their tendency to memorize and leak sensitive information. However, Multi-modal Large Language Models (MLLMs), which process both text and images, introduce unique privacy challenges that remain underexplored. Compared to text-only models, MLLMs can extract and expose sensitive information embedded in images, posing new privacy risks. We reveal that some MLLMs are susceptible to privacy breaches, leaking sensitive data embedded in images or stored in memory. Specifically, in this paper, we (1) introduce MM-Privacy, a comprehensive dataset designed to assess privacy risks across various multi-modal tasks and scenarios, where we define Disclosure Risks and Retention Risks. (2) systematically evaluate different MLLMs using MM-Privacy and demonstrate how models leak sensitive data across various tasks, and (3) provide additional insights into the role of task inconsistency in privacy risks, emphasizing the urgent need for mitigation strategies. Our findings highlight privacy concerns in MLLMs, underscoring the necessity of safeguards to prevent data exposure. Part of our dataset and code can be found here.

DAM: Dynamic Attention Mask for Long-Context Large Language Model Inference Acceleration

Hanzhi Zhang, Heng Fan, Kewei Sha, Yan Huang, Yunhe Feng

Long-context understanding is crucial for many NLP applications, yet transformers struggle with efficiency due to the quadratic complexity of self-attention. Sparse attention methods alleviate this cost but often impose static, predefined masks, failing to capture heterogeneous attention patterns. This results in suboptimal token interactions, limiting adaptability and retrieval accuracy in long-sequence tasks. This work introduces a dynamic sparse attention mechanism that assigns adaptive masks at the attention-map level, preserving heterogeneous patterns across layers and heads. Unlike existing approaches, our method eliminates the need for fine-tuning and predefined mask structures while maintaining computational efficiency. By learning context-aware attention structures, it achieves high alignment with full-attention models, ensuring minimal performance degradation while reducing memory and compute overhead. This approach provides a scalable alternative to full attention, enabling the practical deployment of large-scale Large Langue Models (LLMs) without sacrificing retrieval performance.

Arbiters of Ambivalence: Challenges of using LLMs in No-Consensus tasks

Bhaktipriya Radharapu, Manon Revel, Megan Ung, Sebastian Ruder, Adina Williams

The increasing use of LLMs as substitutes for humans in "aligning" LLMs has raised questions about their ability to replicate human judgments and preferences, especially in ambivalent scenarios where humans disagree. This study examines the biases and limitations of LLMs in three roles: answer generator, judge, and debater. These roles loosely correspond to previously described alignment frameworks: preference alignment (judge) and scalable oversight (debater), with the answer generator reflecting the typical setting with user interactions. We develop a ``no-consensus" benchmark by curating examples that encompass a variety of a priori ambivalent scenarios, each presenting two possible stances. Our results show that while LLMs can provide nuanced assessments when generating open-ended answers, they tend to take a stance on no-consensus topics when employed as judges or debaters. These findings underscore the necessity for more sophisticated methods for aligning LLMs without human oversight, highlighting that LLMs cannot fully capture human non-agreement even on topics where humans themselves are divided.

MM-R^3: On (In-)Consistency of Vision-Language Models (VLMs)

Shih-Han Chou, Shivam Chandhok, Jim Little, Leonid Sigal

With the advent of LLMs and variants, a flurry of research has emerged, analyzing the performance of such models across an array of tasks. While most studies focus on evaluating the capabilities of state-of-the-art (SoTA) Vision Language Models (VLMs) through task accuracy (\textit{e.g.}, visual question answering, grounding), our work explores the related but complementary aspect of \textit{consistency} – the ability of a VLM to produce semantically similar or identical responses to semantically similar queries. We note that consistency is a fundamental prerequisite (necessary but not sufficient condition) for robustness and trust in VLMs. Armed with this perspective, we propose the \textit{MM-R^3 benchmark}, which allows us to analyze performance, in terms of consistency and accuracy, of SoTA VLMs on three tasks: Question Rephrasing, Image Restyling, and Context Reasoning. Our analysis reveals that consistency does not always align with accuracy, indicating that models with higher accuracy are not necessarily more consistent, and vice versa. Furthermore, we propose a simple yet effective mitigation strategy in the form of an adapter module trained to minimize inconsistency across prompts. With our proposed strategy, we are able to achieve absolute improvements of 5.7% and 12.5%, on average on widely used VLMs such as BLIP-2 and LLaVa 1.5M in terms of consistency over their existing counterparts.

Spectral Insights into Data-Oblivious Critical Layers in Large Language Models

Xuyuan Liu, Lei Hsiung, Yaoqing Yang, Yujun Yan

Understanding how feature representations evolve across layers in large language models (LLMs) is crucial for enhancing their interpretability and robustness. Recent research has explored this by identifying critical layers associated with specific functions or behaviors in LLMs. However, prior work has primarily identified critical layers through data-dependent analyses of fine-tuned models, limiting their applicability to post-hoc scenarios. In contrast, we propose a data-oblivious method to pinpoint intrinsic critical layers in pre-fine-tuned LLMs by analyzing representation dynamics using Centered Kernel Alignment (CKA). We demonstrate that layers exhibiting significant changes in CKA correspond to those undergoing the most substantial modifications during fine-tuning, and we show that this phenomenon remains consistent across tasks for a given architecture. Furthermore, our novel spectral analysis reveals that shifts in layer representations are driven by changes in their top principal components, which govern semantic transformations from rationales to conclusions. Notably, freezing these layers during fine-tuning reduces adversarial attack success rates by 40%, enabling efficient, minimally invasive defenses. Our findings suggest that critical layers are an intrinsic property of pre-fine-tuned models, providing a unified framework for understanding LLM layer dynamics and facilitating resource-efficient adaptation strategies.

AdaV: Adaptive Text-visual Redirection for Vision-Language Models

Jiayi Han, Liang Du, Yiwen Wu, Guanming Liang, Xiangguo Zhou, Weibo Zheng, Donghong Han, Zixun Sun

The success of Vision-Language Models (VLMs) often relies on high-resolution schemes that preserve image details, while these approaches also generate an excess of visual tokens, leading to a substantial decrease in model efficiency. A typical VLM includes a visual encoder, a text encoder, and an LLM. Recent studies suggest pruning visual tokens based on visual and textual priors to accelerate VLMs without additional training costs. However, these methods often overlook prompt semantics or suffer from biased self-attention in the LLM. Inspired by the efficient mechanisms of the human brain for multimodal understanding, we introduce AdaV, a novel training-free visual token pruning method. By emulating the neural pathways that preprocess visual and auditory information before the reasoning stage, we shift text-guided visual attention redirection to the pre-LLM stage, which reduces biased token pruning and enhances model robustness with a limited visual token budget. A Self-adaptive Cross-modality Attention Redirection (SCAR) module is further proposed that effectively merges and redirects visual attention with text-to-image attention. Extensive experiments on seven challenging benchmarks demonstrate that our AdaV achieves SOTA performance in training-free VLM acceleration and can be plug-and-play on various VLMs. We plan to open-source the code upon publication.

SuLoRA: Subspace Low-Rank Adaptation for Parameter-Efficient Fine-Tuning

Chenhao Ding, Jiangyang Li, SongLin Dong, Xinyuan Gao, Yuhang He, Yihong Gong

As the scale of large language models (LLMs) grows and natural language tasks become increasingly diverse, Parameter-Efficient Fine-Tuning (PEFT) has become the standard paradigm for fine-tuning LLMs. Among PEFT methods, LoRA is widely adopted for not introducing additional inference overhead. However, existing LoRA's shared parameter space paradigm introduces parameter interference, leading to a gap in generalization performance for specific tasks compared to full fine-tuning. To address this issue, we propose a parameter-separated low-rank adapter, called Subspace Low-Rank Adaptation (SuLoRA). The core idea of SuLoRA is to account for task differences by decomposing LoRA's parameter matrix into multiple independent subspaces and assigning them differentially to distinct tasks. This prevents interference across tasks and enhances the effectiveness of low-rank adaptation. Additionally, SuLoRA achieves higher rank expansion by freezing the A matrix, further improving generalization capability. We conduct extensive experiments on various NLP tasks, demonstrating that SuLoRA significantly outperforms LoRA in trainable parameter efficiency and overall model performance. Furthermore, we validate SuLoRA's effectiveness in domain generalization and multi-modal tasks, showcasing its strong generalization ability.

Context-Aware Hierarchical Merging for Long Document Summarization

Litu Ou, Mirella Lapata

Hierarchical Merging is a technique commonly used to summarize very long texts (>100K tokens) by breaking down the input into smaller sections, summarizing those sections individually, and then merging or combining those summaries into a final coherent summary. Although it helps address the limitations of large language models (LLMs) with fixed input length constraints, the recursive merging process can amplify LLM hallucinations, increasing the risk of factual inaccuracies. In this paper, we seek to mitigate hallucinations by enriching hierarchical merging with context from the source document. Specifically, we propose different approaches to contextual augmentation ranging from *replacing* intermediate summaries with relevant input context, to *refining* them while using the context as supporting evidence, and *aligning* them implicitly (via citations) to the input. Experimental results on datasets representing legal and narrative domains show that contextual augmentation consistently outperforms zero-shot and hierarchical merging baselines for the Llama 3.1 model family. Our analysis further reveals that refinement methods tend to perform best when paired with extractive summarization for identifying relevant input.

FGDGNN: Fine-Grained Dynamic Graph Neural Network for Rumor Detection on Social Media

Mei Guo, Chen Chen, Chunyan Hou, Yike Wu, Xiaojie Yuan

Detecting rumors on social media has become a crucial issue. Propagation structure-based methods have recently attracted increasing attention. When the propagation structure is represented by the dynamic graph, temporal information is considered. However, existing rumor detection models using dynamic graph typically focus only on coarse-grained temporal information and ignore the fine-grained temporal dynamics within individual snapshots and across snapshots. In this paper, we propose a novel Fine-Grained Dynamic Graph Neural Network (FGDGNN) model, which can incorporate the fine-grained temporal information of dynamic propagation graph in the intra-snapshot and dynamic embedding update mechanism in the inter-snapshots into a unified framework for rumor detection. Specifically, we firstly construct the edge-weighed propagation graph and the edge-aware graph isomorphism network is proposed. To get fine-grained temporal representations across snapshots, we propose an embedding transformation layer to update node embedding. Finally, we integrate the temporal information in the inter-snapshots at the graph level to enhance the effectiveness of the proposed model. Extensive experiments conducted on three public real-world datasets demonstrate that our FGDGNN model achieves significant improvements compared with the state-of-the-art baselines.

CA-GAR: Context-Aware Alignment of LLM Generation for Document Retrieval

Heng Yu, Junfeng Kang, Rui Li, Qi Liu, Liyang He, Zhenya Huang, Shuanghong Shen, Junyu Lu

Information retrieval has evolved from traditional sparse and dense retrieval methods to approaches driven by large language models (LLMs). Recent techniques, such as Generation-Augmented Retrieval (GAR) and Generative Document Retrieval (GDR), leverage LLMs to enhance retrieval but face key challenges: GAR's generated content may not always align with the target document corpus, while GDR limits the generative capacity of LLMs by constraining outputs to predefined document identifiers. To address these issues, we propose Context-Aware Generation-Augmented Retrieval (CA-GAR), which enhances LLMs by integrating corpus information into their generation process. CA-GAR optimizes token selection by incorporating relevant document information and leverages a Distribution Alignment Strategy to extract corpus information using a lexicon-based approach. Experimental evaluations on seven tasks from the BEIR benchmark and four non-English languages from Mr.TyDi demonstrate that CA-GAR outperforms existing methods.

Expectation Confirmation Preference Optimization for Multi-Turn Conversational Recommendation Agent

Xueyang Feng, Jingsen Zhang, Jiakai Tang, Wei Li, Guohao Cai, Xu Chen, Quanyu Dai, Yue Zhu, Zhenhua Dong

Recent advancements in Large Language Models (LLMs) have significantly propelled the development of Conversational Recommendation Agents (CRAs). However, these agents often generate short-sighted responses that fail to sustain user guidance and meet expectations. Although preference optimization has proven effective in aligning LLMs with user expectations, it remains costly and performs poorly in multi-turn dialogue. To address this challenge, we introduce a novel multi-turn preference optimization (MTPO) paradigm **ECPO**, which leverages Expectation Confirmation Theory to explicitly model the evolution of user satisfaction throughout multi-turn dialogues, uncovering the underlying causes of dissatisfaction. These causes can be utilized to support targeted optimization of unsatisfactory responses, thereby achieving turn-level preference optimization. ECPO eliminates the significant sampling overhead of existing MTPO methods while ensuring the optimization process drives meaningful improvements. To support ECPO, we also introduce an LLM-based user simulator, **AILO**, to simulate user feedback and expectation confirmation during conversational recommendations. Experimental results show that ECPO significantly enhances CRA's interaction capabilities, offering notable improvements in both efficiency and effectiveness over existing MTPO methods.

RASD: Retrieval-Augmented Speculative Decoding

Guofeng Quan, Wenfeng Feng, Chuzhan Hao, Guochao Jiang, Yuewei Zhang, Hao Henry Wang

Speculative decoding accelerates inference in large language models (LLMs) by generating draft tokens for target model verification. Current approaches for obtaining draft tokens rely on lightweight draft models or additional model structures to generate draft tokens and retrieve context from databases. Due to the draft model's small size and limited training data, model-based speculative decoding frequently becomes less effective in out-of-domain scenarios. Additionally, the time cost of the drafting phase results in a low upper limit on acceptance length during the verification step, limiting overall efficiency. This paper proposes RASD (Retrieval-Augmented Speculative Decoding), which adopts retrieval methods to enhance model-based speculative decoding. We introduce tree pruning and tree fusion to achieve this. Specifically, we develop a pruning method based on the draft model's probability distribution to construct the optimal retrieval tree. Second, we employ the longest prefix matching algorithm to merge the tree generated by the draft model with the retrieval tree, resulting in a unified tree for verification. Experimental results demonstrate that RASD achieves state-of-the-art inference acceleration across tasks such as DocQA, Summary, Code, and In-Domain QA. Moreover, RASD exhibits strong scalability, seamlessly integrating with various speculative decoding approaches, including both generation-based and retrieval-based methods.

Reefknot: A Comprehensive Benchmark for Relation Hallucination Evaluation, Analysis and Mitigation in Multimodal Large Language Models

Kening Zheng, Junkai Chen, Yibo Yan, Xin Zou, Huiyu Zhou, Xuming Hu

Hallucination issues continue to affect multimodal large language models (MLLMs), with existing research mainly addressing object-level or attribute-level hallucinations, neglecting the more complex relation hallucinations that require advanced reasoning. Current benchmarks for relation hallucinations lack detailed evaluation and effective mitigation, and their datasets often suffer from biases due to systematic annotation processes. To address these challenges, we introduce Reefknot, a comprehensive benchmark targeting relation hallucinations, comprising over 20,000 real-world samples. We provide a systematic definition of relation hallucinations, integrating perceptive and cognitive perspectives, and construct a relation-based corpus using the Visual Genome scene graph dataset. Our comparative evaluation reveals significant limitations in current MLLMs' ability to handle relation hallucinations. Additionally, we propose a novel confidence-based mitigation strategy, which reduces the hallucination rate by an average of 9.75% across three datasets, including Reefknot. Our work offers valuable insights for achieving trustworthy multimodal intelligence. The dataset and code are released at https://github.com/JackChen-seu/Reefknot.

Explorer: Scaling Exploration-driven Web Trajectory Synthesis for Multimodal Web Agents

Vardaan Pahuja, Yadong Lu, Corby Rosset, Boyu Gou, Arindam Mitra, Spencer Whitehead, Yu Su, Ahmed Hassan Awadallah

Recent success in large multimodal models (LMMs) has sparked promising applications of agents capable of autonomously completing complex web tasks. While open-source LMM agents have made significant advances in offline evaluation benchmarks, their performance still falls substantially short of human-level capabilities in more realistic online settings. A key bottleneck is the lack of diverse and large-scale trajectory-level datasets across various domains, which are expensive to collect. In this paper, we address this challenge by developing a scalable recipe to synthesize the largest and most diverse trajectory-level dataset to date, containing over 94K successful multimodal web trajectories, spanning 49K unique URLs, 720K screenshots, and 33M web elements. In particular, we leverage extensive web exploration and refinement to obtain diverse task intents. The average cost is 28 cents per successful trajectory, making it affordable to a wide range of users in the community. Leveraging this dataset, we train Explorer, a multimodal web agent, and demonstrate strong performance on both offline and online web agent benchmarks such as Mind2Web-Live, Multimodal-Mind2Web, and MiniWob++. Additionally, our experiments highlight data scaling as a key driver for improving web agent capabilities. We hope this study makes state-of-the-art LMM-based agent research at a larger scale more accessible.

EssayJudge: A Multi-Granular Benchmark for Assessing Automated Essay Scoring Capabilities of Multimodal Large Language Models

Jiamin Su, Yibo Yan, Fangteng FU, Zhang Han, Jingheng Ye, Xiang Liu, Jiahao Huo, Huiyu Zhou, Xuming Hu

Automated Essay Scoring (AES) plays a crucial role in educational assessment by providing scalable and consistent evaluations of writing tasks. However, traditional AES systems face three major challenges: (i) reliance on handcrafted features that limit generalizability, (ii) difficulty in capturing fine-grained traits like coherence and argumentation, and (iii) inability to handle multimodal contexts. In the era of Multimodal Large Language Models (MLLMs), we propose **EssayJudge**, the **first multimodal benchmark to evaluate AES capabilities across lexical-, sentence-, and discourse-level traits**. By leveraging MLLMs' strengths in trait-specific scoring and multimodal context understanding, EssayJudge aims to offer precise, context-rich evaluations without manual feature engineering, addressing longstanding AES limitations. Our experiments with 18 representative MLLMs reveal gaps in AES performance compared to human evaluation, particularly in discourse-level traits, highlighting the need for further advancements in MLLM-based AES research. Our dataset and code will be available upon acceptance.

Streamlining the Collaborative Chain of Models into A Single Forward Pass in Generation-Based Tasks

Yuanjie Lyu, Chao Zhang, Yuhao Chen, Yong Chen, Tong Xu

In Retrieval-Augmented Generation (RAG) and agent-based frameworks, the "Chain of Models" approach is widely used, where multiple specialized models work sequentially on distinct sub-tasks. This approach is effective but increases resource demands as each model must be deployed separately. Recent advancements attempt to address this by applying prompt tuning, which allows a shared base model to adapt to multiple tasks with minimal parameter changes. However, a key challenge remains: intermediate outputs, passed between models as plain text, require recomputation of hidden states (i.e., Key and Value (KV) states in Transformers) during inference. In this paper, we introduce FTHSS, a novel prompt-tuning method that enables models to share KV hidden states, eliminating redundant forward passes and reducing KV cache storage. By modifying input and attention masks during training, FTHSS allows models to effectively utilize KV hidden states from prior models in both single- and multi-round scenarios. Empirical results on four tasks show that FTHSS matches the performance of traditional model chains while improving inference efficiency.

Firm or Fickle? Evaluating Large Language Models Consistency in Sequential Interactions

Yubo Li, Yidi Miao, Xueying Ding, Ramayya Krishnan, Rema Padman

Large Language Models (LLMs) have shown remarkable capabilities across various tasks, but their deployment in high-stake domains requires consistent performance across multiple interaction rounds. This paper introduces a comprehensive framework for evaluating and improving LLM response consistency, making three key contributions. First, we propose a novel Position-Weighted Consistency (PWC) score that captures both the importance of early-stage stability and recovery patterns in multi-turn interactions. Second, we present a carefully curated benchmark dataset spanning diverse domains and difficulty levels, specifically designed to evaluate LLM consistency under various challenging follow-up scenarios. Third, we introduce Confidence-Aware Response Generation (CARG), a framework that significantly improves response stability by incorporating model confidence signals into the generation process. Empirical results demonstrate that CARG significantly improves response stability without sacrificing accuracy, underscoring its potential for reliable LLM deployment in critical applications.

Logical DA: Enhancing Data Augmentation for Logical Reasoning via a Multi-Agent System

Haoqi Zheng, DongWang, Silin Yang, Yunpeng Qi, Ruochun Jin

Recent advancements in large language models (LLMs) have highlighted the importance of improving their reasoning capabilities. A critical challenge lies in the scarcity of high-quality reasoning data—characterized by diversity and rich supervisory signals—necessary for robust model training. While data augmentation (DA) methods have been leveraged to mitigate this scarcity, prevailing approaches often introduce noise and exhibit logical inconsistencies, thereby diminishing their utility for complex reasoning tasks. Moreover, existing DA paradigms predominantly isolate data synthesis from label validation, failing to unify these complementary processes within a cohesive architecture. To address these limitations, we introduce Logical DA, a multi-agent framework for enhancing reasoning-focused data augmentation in few-shot learning scenarios. Our system includes four agents operating through two synergistic phases: (1) diverse data generation, and (2) label verification. The system incorporates a reflection mechanism to continuously improve data quality by leveraging feedback from logical validation. We demonstrate the effectiveness of Logical DA through experiments on various tasks and datasets, achieving the highest average improvement in task accuracy in both fine-tuning and in-context learning paradigms, with an average improvement of 7.61\% when applied to fine-tuning.

KodCode: A Diverse, Challenging, and Verifiable Synthetic Dataset for Coding

Zhangchen Xu, Yang Liu, Yueqin Yin, Mingyuan Zhou, Radha Poovendran

We introduce KodCode, a synthetic dataset that addresses the persistent challenge of acquiring high-quality, verifiable training data across diverse difficulties and domains for training Large Language Models for coding. Existing code-focused resources typically fail to ensure either the breadth of coverage (e.g., spanning simple coding tasks to advanced algorithmic problems) or verifiable correctness (e.g., unit tests). In contrast, KodCode comprises question–solution–test triplets that are systematically validated via a self-verification procedure. Our pipeline begins by synthesizing a broad range of coding questions, then generates solutions and test cases with additional attempts allocated to challenging problems. Finally, post-training data synthesis is done by rewriting questions into diverse formats and generating responses under a test-based reject sampling procedure from a reasoning model (DeepSeek R1). This pipeline yields a large-scale, robust and diverse coding dataset. It is suitable for supervised fine-tuning and the paired unit tests also provide great potential for RL tuning. Fine-tuning experiments on coding benchmarks (HumanEval(+), MBPP(+), BigCodeBench, and LiveCodeBench) demonstrate that KodCode-tuned models achieve state-of-the-art performance, surpassing models like Qwen2.5-Coder-32B-Instruct and DeepSeek-R1-Distill-Llama-70B.

MMUnlearner: Reformulating Multimodal Machine Unlearning in the Era of Multimodal Large Language Models

Jiahao Huo, Yibo Yan, Xu Zheng, Yuanhuiyi Lyu, Xin Zou, Zhihua Wei, Xuming Hu

Recent progress in Machine Unlearning (MU) has introduced solutions for the selective removal of private or sensitive information encoded within deep neural networks. Nonetheless, MU for Multimodal Large Language Models (MLLMs) remains in its nascent phase. Therefore, we propose to **reformulate the task of multimodal MU in the era of MLLMs**, which aims to erase only the visual patterns associated with a given entity while preserving the corresponding textual knowledge encoded within the original parameters of the language model backbone. Furthermore, we **develop a novel geometry-constrained gradient ascent method MMUnlearner**. It updates the weights of MLLMs with a weight saliency map jointly restricted by the remaining concepts and textual knowledge during unlearning, thereby preserving parameters essential for non-target knowledge. Extensive experiments demonstrate that MMUnlearner surpasses baselines that finetuning MLLMs with VQA data directly through Gradient Ascent (GA) or Negative Preference Optimization (NPO), across all evaluation dimensions. Our code will be released upon acceptance.

DiffSkip: Differential Layer Skipping in Large Language Models

Xuan Luo, Weizhi Wang, Xifeng Yan

Existing Large Language Models (LLMs) enforce uniform computation across all tokens. We analyze the correlation between the input-output difference of self-attention block and Feed-Forward Network (FFN) within the same transformer layer, and find that these two differential vectors are highly correlated. Thus, we propose to dynamically skip the FFN blocks based on the self-attention difference and introduce Diffential Layer Skipping (DiffSkip) to show that LLMs are inherently dynamic-depth models, capable of adjusting computational depth when generating different tokens. DiffSkip employs a lightweight router module to dynamically skip a set of FFN blocks in LLMs and only requires efficient fine-tuning while keeping the whole LLM frozen. Experimental results demonstrate that DiffSkip effectively enables dynamic FFN skipping in decoder-only language models, even in continuous token generation tasks where many layer-skipping methods struggle.

LongDPO: Unlock Better Long-form Generation Abilities for LLMs via Critique-augmented Stepwise Information

Bowen Ping, Jiali Zeng, Fandong Meng, Shuo Wang, Jie Zhou, Shanghang Zhang

Recent advancements in large language models (LLMs) have markedly improved their capacity to handle long text inputs; however, current models, including GPT-4o, still exhibit unsatisfactory performance in long-form generation. Generating high-quality long-form content still remains a significant challenge. In this paper, we present LongDPO, a novel approach designed to enhance long-form text generation through step-level supervision. By leveraging Monte Carlo Tree Search (MCTS) to collect stepwise preference pairs and employing a global memory pool to maintain factual accuracy, LongDPO effectively mitigates issues such as inconsistencies that are prevalent in long-context LLMs. Furthermore, we integrate critique-augmented generation to refine the selected preference pairs. Following the collection of stepwise preference pairs, we apply stepwise preference learning for fine-grained optimization. Experimental results demonstrate that our method enhances performance on long-form generation benchmarks (e.g. LongBench-Write) while maintaining nearly lossless performance on several general benchmarks.

FRAME: Feedback-Refined Agent Methodology for Enhancing Medical Research Insights

Chengzhang Yu, Yiming Zhang, Zhixin Liu, Zenghui Ding, Yining Sun, Zhanpeng Jin

The automation of scientific research through large language models (LLMs) presents significant opportunities but faces critical challenges in knowledge synthesis and quality assurance. We introduce Feedback-Refined Agent Methodology (FRAME), a novel framework that enhances medical paper generation through iterative refinement and structured feedback. Our approach comprises three key innovations: (1) A structured dataset construction method that decomposes 4,287 medical papers into essential research components through iterative refinement; (2) A tripartite architecture integrating Generator, Evaluator, and Reflector agents that progressively improve content quality through metric-driven feedback; and (3) A comprehensive evaluation framework that combines statistical metrics with human-grounded benchmarks. Experimental results demonstrate FRAME's effectiveness, achieving significant improvements over conventional approaches across multiple models (9.91% average gain with DeepSeek V3, comparable improvements with GPT-4o Mini) and evaluation dimensions. Human evaluation confirms that FRAME-generated papers achieve quality comparable to human-authored works, with particular strength in synthesizing future research directions. The results demonstrated our work could efficiently assist medical research by building a robust foundation for automated medical research paper generation while maintaining rigorous academic standards.

CodePRM: Execution Feedback-enhanced Process Reward Model for Code Generation

Qingyao Li, Xinyi Dai, Xiangyang Li, Weinan Zhang, Yasheng Wang, Ruiming Tang, Yong Yu

Code generation is a critical reasoning task for large language models (LLMs). Recent advancements have focused on optimizing the thought process of code generation, achieving significant improvements. However, such thought process lacks effective process supervision, making it hard to optimize the thoughts. Although Process Reward Models (PRMs) have been widely established in mathematical reasoning, building a code PRM is still not trivial for the gap between thoughts to code. In this paper, we propose CodePRM, a novel approach that leverages the code execution feedback to build a code PRM. Specifically, we first collect a large dataset of thought traces, where each thought step is labeled with their derived code' pass rates, accompanied by the corresponding code snippets, and execution feedback. During training, we train a PRM to take both the reasoning process and code execution feedback as input to score individual thought steps, enabling it to leverage code execution results to distinguish between high-quality and low-quality thought steps. Finally, to use the PRM during inference, we develop a Generate-Verify-Refine (GVR) pipeline where the CodePRM serves as a process verifier to dynamically identify and correct errors in the thought process during code search. Experimental results demonstrate that CodePRM with the inference algorithm outperforms strong baselines, significantly enhancing code generation performance. Further analysis reveals the key factors for building a code PRM.

Leveraging LLMs for Bangla Grammar Error Correction: Error Categorization, Synthetic Data, and Model Evaluation

Pramit Bhattacharyya, Arnab Bhattacharya

Large Language Models (LLMs) perform exceedingly well in Natural Language Understanding (NLU) tasks for many languages including English. However, despite being the fifth most-spoken language globally, Grammatical Error Correction (GEC) in Bangla remains underdeveloped. In this work, we investigate how LLMs can be leveraged for improving Bangla GEC. For that, we first do an extensive categorization of 12 error classes in Bangla, and take a survey of native Bangla speakers to collect real-world errors. We next devise a rule-based noise injection method to create grammatically incorrect sentences corresponding to correct ones. The Vaiyākaraṇa dataset, thus created, consists of 567,422 sentences of which 227,119 are erroneous. This dataset is then used to instruction-tune LLMs for the task of GEC in Bangla. Evaluations show that instruction-tuning with Vaiyākaraṇa improves GEC performance of LLMs by 3-7 percentage points as compared to zero-shot setting, and makes them achieve human-like performance in error identification. Humans, though, remain superior in grammatical error correction.

REALM: A Dataset of Real-World LLM Use Cases

Jingwen Cheng, Kshitish Ghate, Wenyue Hua, William Yang Wang, Hong Shen, Fei Fang

Large Language Models (LLMs), such as the GPT series, have driven significant industrial applications, leading to economic and societal transformations. However, a comprehensive understanding of their real-world applications remains limited. To address this, we introduce REALM, a dataset of over 94,000 LLM use cases collected from Reddit and news articles. REALM captures two key dimensions: the diverse applications of LLMs and the demographics of their users. It categorizes LLM applications and explores how users' occupations relate to the types of applications they use. By integrating real-world data, REALM offers insights into LLM adoption across different domains, providing a foundation for future research on their evolving societal roles.

Retrieval-Augmented Process Reward Model for Generalizable Mathematical Reasoning

Jiachen Zhu, Congmin Zheng, Jianghao Lin, Kounianhua Du, Ying Wen, Yong Yu, Jun Wang, Weinan Zhang

While large language models (LLMs) have significantly advanced mathematical reasoning, Process Reward Models (PRMs) have been developed to evaluate the logical validity of reasoning steps. However, PRMs still struggle with out-of-distribution (OOD) challenges. This paper identifies the OOD issues including step OOD, arising from differences in reasoning patterns across model types and sizes, and question OOD, due to dataset shifts between training and real-world problems. To address these issues, we introduce Retrieval-Augmented Process Reward Model (RetrievalPRM), a novel framework designed to tackle these OOD issues. By utilizing a two-stage retrieval-enhanced mechanism, RetrievalPRM retrieves semantically similar questions and steps for PRM as a warmup to stimulate its potential to judge target steps, improving generalization and reasoning consistency across different models and problem types. Our extensive experiments demonstrate that RetrievalPRM outperforms existing baselines across multiple real-world datasets. Our open-source contributions include a retrieval-enhanced dataset, a tuning framework for PRM training, and the RetreivalPRM model, establishing a new standard for PRM performance.

XFinBench: Benchmarking LLMs in Complex Financial Problem Solving and Reasoning

Zhihan Zhang, Yixin Cao, Lizi Liao

Solving financial problems demands complex reasoning, multimodal data processing, and a broad technical understanding, presenting unique challenges for current large language models (LLMs). We introduce **XFinBench**, a novel benchmark with 4,235 examples designed to evaluate LLM's ability in solving comple**X**, knowledge-intensive **Fin**ancial problems across diverse graduate-level finance topics with multi-modal context. We identify five core capabilities of LLMs using XFinBench, i.e., _terminology understanding_, _temporal reasoning_, _future forecasting_, _scenario planning_, and _numerical modelling_. Upon XFinBench, we conduct extensive experiments on 18 leading models. The result shows that o1 is the best-performing text-only model with an overall accuracy of 67.3%, but still lags significantly behind human experts with 12.5%, especially in temporal reasoning and scenario planning capabilities. We further construct a knowledge bank with 3,032 finance terms for knowledge augmentation analysis, and find that relevant knowledge to the question only brings consistent accuracy improvements to small open-source model. Additionally, our error analysis reveals that rounding errors during calculation and blindness to position and intersection of curves in the image are two primary issues leading to model's poor performance in calculating and visual-context questions, respectively.

Chain of Attack: Hide Your Intention through Multi-Turn Interrogation

Xikang Yang, biyu zhou, Xuehai Tang, Jizhong Han, Songlin Hu

The latent knowledge of large language models (LLMs) contains harmful or unethical content, which introduces significant security risks upon their widespread deployment. Conducting jailbreak attacks on LLMs can proactively identify vulnerabilities to enhance their security measures. However, previous jailbreak attacks primarily focus on single-turn dialogue scenarios, leaving vulnerabilities in multi-turn dialogue contexts inadequately explored. This paper investigates the resilience of black-box LLMs in multi-turn jailbreak attack scenarios from a novel interrogation perspective. We propose an optimal interrogation principle to conceal the jailbreak intent and introduce a multi-turn attack chain generation strategy called CoA. By employing two effective interrogation strategies tailored for LLMs, coupled with an interrogation history record management mechanis, it achieves a significant optimization of the attack process. Our approach enables the iterative generation of attack chains, offering a powerful tool for LLM red team testing. Experimental results demonstrate that LLMs exhibit insufficient resistance under multi-turn interrogation, with our method shows more advantages(ASR, 83% vs 64%). This work offers new insights into improving the safety of LLMs.

Learning to Play Like Humans: A Framework for LLM Adaptation in Interactive Fiction Games

Jinming Zhang, Yunfei Long

Interactive Fiction games (IF games) are where players interact through natural language commands. While recent advances in Artificial Intelligence agents have reignited interest in IF games as a domain for studying decision-making, existing approaches prioritize task-specific performance metrics over human-like comprehension of narrative context and gameplay logic. This work presents a cognitively inspired framework that guides Large Language Models (LLMs) to learn and play IF games systematically. Our proposed **L**earning to **P**lay **L**ike **H**umans (LPLH) framework integrates three key components: (1) structured map building to capture spatial and narrative relationships, (2) action learning to identify context-appropriate commands, and (3) feedback-driven experience analysis to refine decision-making over time. By aligning LLMs-based agents' behavior with narrative intent and commonsense constraints, LPLH moves beyond purely exploratory strategies to deliver more interpretable, human-like performance. Crucially, this approach draws on cognitive science principles to more closely simulate how human players read, interpret, and respond within narrative worlds. As a result, LPLH reframes the IF games challenge as a learning problem for LLMs-based agents, offering a new path toward robust, context-aware gameplay in complex text-based environments.

ESF: Efficient Sensitive Fingerprinting for Black-Box Tamper Detection of Large Language Models

Xiaofan Bai, Pingyi Hu, Xiaojing Ma, Bin Benjamin Zhu, Linchen Yu, Dongmei Zhang, Qi Zhang

The rapid adoption of large language models (LLMs) in diverse applications has intensified concerns over their security and integrity, especially in cloud environments where internal model parameters are inaccessible to users. Traditional tamper detection methods, designed for deterministic classification models, fail to address the output randomness and massive parameter spaces characteristic of LLMs. In this paper, we introduce \emph{Efficient Sensitive Fingerprinting (ESF)}, the first fingerprinting method tailored for black-box tamper detection of LLMs. ESF generates fingerprint samples by optimizing output sensitivity at selected detection token positions and leverages \emph{Randomness-Set Consistency Checking (RSCC)} to accommodate inherent output randomness. Furthermore, a novel \emph{Max Coverage Strategy (MCS)} is proposed to select an optimal set of fingerprint samples that maximizes joint sensitivity to tampering. Grounded in a rigorous theoretical framework, ESF is both computationally efficient and scalable to large models. Extensive experiments across state-of-the-art LLMs demonstrate that ESF reliably detects tampering, such as fine-tuning, model compression, and backdoor injection, with a detection rate exceeding 99.2\%, thereby offering a robust solution for securing cloud-based AI systems.

ProBench: Judging Multimodal Foundation Models on Open-ended Multi-domain Expert Tasks

Yan Yang, Dongxu Li, Haoning Wu, Bei Chen, Liu Liu, Liyuan Pan, Junnan Li

Solving expert-level multimodal tasks is a key milestone in general intelligence. As the capabilities of multimodal large language models (MLLMs) continue to evolve, evaluation of frontier multimodal intelligence becomes necessary yet challenging. In this work, we introduce ProBench, a benchmark of open-ended user queries encapsulating professional expertise and advanced reasoning. ProBench consists of 4,000 high-quality samples independently collected from professionals based on their productivity demands. It spans across 10 fields and 56 sub-fields, including science, arts, humanities, coding, mathematics, and creative writing. Experimentally, we evaluate and compare 24 latest models using MLLM-as-a-Judge. Our results reveal that although the best open-source models rival the proprietary ones, they all face significant challenges in visual perception, textual understanding, domain knowledge, and advanced reasoning. Our benchmark is publicly accessible at .

Investigating and Scaling up Code-Switching for Multilingual Language Model Pre-Training

Zhijun Wang, Jiahuan Li, Hao zhou, Rongxiang Weng, Jingang Wang, Xin Huang, Xue Han, Junlan Feng, Chao Deng, Shujian Huang

Large language models (LLMs) exhibit remarkable multilingual capabilities despite the extreme language imbalance in the pre-training data. In this paper, we closely examine the reasons behind this phenomenon, focusing on the pre-training corpus. We find that the existence of code-switching, alternating between different languages within a context, is key to multilingual capabilities. We conduct an analysis to investigate code-switching in the pre-training corpus, examining its presence and categorizing it into four types within two quadrants. We then assess its impact on multilingual performance. These types of code-switching data are unbalanced in proportions and demonstrate different effects on facilitating language transfer. To better explore the power of code-switching for language alignment during pre-training, we investigate the strategy of synthetic code-switching. We continuously scale up the synthetic code-switching data and observe remarkable improvements in both benchmarks and representation space. Extensive experiments indicate that incorporating synthetic code-switching data enables better language alignment and generalizes well to high, medium, and low-resource languages with pre-training corpora of varying qualities.

Rhetorical Device-Aware Sarcasm Detection with Counterfactual Data Augmentation

Qingqing Hong, Dongyu Zhang, Jiayi Lin, Dapeng Yin, Shuyue Zhu, Junli Wang

Sarcasm is a complex form of sentiment expression, widely used in human daily life. Previous work primarily defines sarcasm as a form of verbal irony, which covers only a subset of real-world sarcastic expressions. However, sarcasm serves multifaceted functions and manifests through various rhetorical devices, such as echoic mentions, rhetorical questions, and hyperbole. To fully capture its complexity, this paper investigates fine-grained sarcasm classification through the lens of rhetorical devices, and introduces \textbf{RedSD}, a \textbf{R}h\textbf{E}torical \textbf{D}evice-Aware \textbf{S}arcasm \textbf{D}ataset with counterfactually augmented data. To construct the dataset, we extract sarcastic dialogues from situation comedies (i.e., sitcoms), and summarize nine rhetorical devices commonly employed in sarcasm. We then propose a rhetorical device-aware counterfactual data generation pipeline facilitated by both Large Language Models (LLMs) and human revision. Additionally, we propose duplex counterfactual augmentation that generates counterfactuals for both sarcastic and non-sarcastic dialogues, to further enhance the scale and diversity of the dataset. Experimental results on the dataset demonstrate that the fine-tuning models show more balanced performance over zero-shot models, including GPT-3.5 and LLaMA 3.1, underscoring the importance of integrating various rhetorical devices in sarcasm detection. Our dataset are avaliable at https://anonymous.4open.science/r/RedSD-742D.

Bridging Robustness and Generalization Against Word Substitution Attacks in NLP via the Growth Bound Matrix Approach

Mohammed Bouri, Adnane Saoud

Despite advancements in Natural Language Processing (NLP), models remain vulnerable to adversarial attacks, such as synonym substitutions. While prior work has focused on improving robustness for feed-forward and convolutional architectures, the robustness of recurrent networks and modern state space models (SSMs), such as S4, remains understudied. These architectures pose unique challenges due to their sequential processing and complex parameter dynamics. In this paper, we introduce a novel regularization technique based on Growth Bound Matrices (GBM) to improve NLP model robustness by reducing the impact of input perturbations on model outputs. We focus on computing the GBM for three architectures: Long Short-Term Memory (LSTM), State Space models (S4), and Convolutional Neural Networks (CNN). Our method aims to (1) enhance resilience against word substitution attacks, (2) improve generalization on clean text, and (3) providing the first systematic analysis of SSM (S4) robustness. Extensive experiments across multiple architectures and benchmark datasets demonstrate that our method improves adversarial robustness by up to (8.8%) over existing baselines. These results highlight the effectiveness of our approach, outperforming several state-of-the-art methods in adversarial defense.

Socratic Style Chain-of-Thoughts Help LLMs to be a Better Reasoner

Jiangbo Pei, Peiyu Liu, Xin Zhao, Aidong Men, Yang Liu

Synthetic data generation has emerged as a promising approach to enhance the reasoning capabilities of large language models. However, existing methods remain hindered by high costs—either through expensive API access or additional intermediate training—and are limited in their ability to generalize across different domains. To address these challenges, we propose a multi-agent debate framework based on the Socratic questioning strategy, abbreviated as SoDa. Distinguished from previous methods that prioritize data quantity, we highlight the wisdom of Socratic questioning in augmenting reasoning quality by deepening the thinking process to encourage exploration and broadening it to motivate self-reflection on each question. Combined with our efficient production pipeline, SoDa enables scaling while maintaining affordable costs. We use SoDa to generate diverse datasets for mathematics and code generation tasks with the Qwen2.5-7B-Instruct model, successfully fine-tuning a range of foundation models, from general-purpose ones to OpenAI o1-like ones. For mathematics, the experimental results show that SoDa outperforms the performance of existing datasets at the same scale, achieving improvements ranging from 1.3% to 13.5%. Remarkably, SoDa with 30K examples even surpasses the ScaleQuest dataset with 1000K samples, demonstrating significant efficiency. Our findings highlight the potential of SoDa as a universal, scalable, and cost-effective method for enhancing reasoning capabilities in large models across domains.

Grounding Task Assistance with Multimodal Cues from a Single Demonstration

Gabriel Herbert Sarch, Balasaravanan Thoravi Kumaravel, Sahithya Ravi, Vibhav Vineet, Andy Wilson

A person's demonstration often serves as a key reference for others learning the same task. However, RGB video, the dominant medium for representing these demonstrations, often fails to capture fine-grained contextual cues such as intent, safety-critical environmental factors, and subtle preferences embedded in human behavior. This sensory gap fundamentally limits the ability of Vision Language Models (VLMs) to reason about why actions occur and how they should adapt to individual users. To address this, we introduce MICA (Multimodal Interactive Contextualized Assistance), a framework that improves conversational agents for task assistance by integrating eye gaze and speech cues. MICA segments demonstrations into meaningful subtasks, extracts keyframes, and generates contextualized captions to improve retrieval-augmented generation. Evaluations on questions derived from real-time chat-assisted task replication show that multimodal cues significantly improve response quality over frame-based retrieval. Notably, gaze cues alone achieves 93% of speech performance, and their combination yields the highest accuracy. Task type determines the effectiveness of implicit (gaze) vs. explicit (speech) cues, underscoring the need for adaptable multimodal models. These results highlight the limitations of frame-based context and demonstrate the value of multimodal signals for real-world AI task assistance.

Dual Debiasing for Noisy In-Context Learning for Text Generation

Siqi Liang, Sumyeong Ahn, Paramveer Dhillon, Jiayu Zhou

In-context learning (ICL) relies heavily on high-quality demonstrations drawn from large annotated corpora. Existing methods often detect noisy annotations by ranking local perplexities, presuming that noisy samples yield higher perplexities than their clean counterparts. However, this assumption breaks down when the noise ratio is high and many demonstrations are flawed. We re-examine the perplexity-based paradigm for text generation under noisy annotations, highlighting two sources of bias in perplexity: the annotation itself and the domain-specific knowledge inherent in large language models (LLMs). To overcome these biases, we introduce a dual-debiasing framework that uses synthesized neighbors to explicitly correct perplexity estimates, yielding a robust \emph{Sample Cleanliness Score}. This metric uncovers absolute sample cleanliness regardless of the corpus's noise level. Extensive experiments demonstrate our method's superior noise-detection capabilities and show that its final ICL performance is comparable to that of a fully clean demonstration corpus. Moreover, our approach remains robust even when noise ratios are extremely high.

A Step Towards Explainable Hate Speech Detection

Happy Khairunnisa Sariyanto, Diclehan Ulucan, Oguzhan Ulucan, Marc Ebner

Recent advancements in deep learning have significantly enhanced the efficiency and accuracy of natural language processing (NLP) tasks. However, these models often require substantial computational resources, which remains a major drawback. Reducing the complexity of deep learning architectures, and exploring simpler yet effective approaches can lead to cost-efficient NLP solutions. This is also a step towards explainable AI, i.e., uncovering how a particular task is carried out. For this analysis, we chose the task of hate speech detection. We address hate speech detection by introducing a model that employs a weighted sum of valence, arousal, and dominance (VAD) scores for classification. To determine the optimal weights and classification strategies, we analyze hate speech and non-hate speech words based on both their individual and summed VAD-values. Our experimental results demonstrate that this straightforward approach can compete with state-of-the-art neural network methods, including GPT-based models, in detecting hate speech.

RoseRAG: Robust Retrieval-augmented Generation with Small-scale LLMs via Margin-aware Preference Optimization

Tianci Liu, Haoxiang Jiang, Tianze Wang, Ran Xu, Yue Yu, Linjun Zhang, Tuo Zhao, Haoyu Wang

Large language models (LLMs) have achieved impressive performance but face high computational costs and latency, limiting their deployment in resource-constrained settings. In contrast, small-scale LLMs (SLMs) are more efficient yet struggle to capture evolving real-world knowledge. Retrieval-augmented generation (RAG) helps by integrating external knowledge, but imperfect retrieval can introduce distracting noise that misleads SLMs. We propose RoseRAG, a robust RAG framework for SLMs via Margin-aware Preference Optimization. RoseRAG employs multi-turn prompting for detailed reasoning, rejection sampling for high-quality explanations, and contrastive preference selection to refine responses by maximizing the likelihood gap between preferred and non-preferred outputs. By integrating these components into a margin-aware optimization process, RoseRAG robustly enhances the accuracy and reliability of SLMs for RAG applications. Extensive experiments on three open-domain question answering benchmarks indicate that our innovative RoseRAG surpasses state-of-the-art baselines significantly.

mRAKL: Multilingual Retrieval-Augmented Knowledge Graph Construction for Low-Resourced Languages

Hellina Hailu Nigatu, Min Li, Maartje Ter Hoeve, Saloni Potdar, Sarah Chasins

Knowledge Graphs represent real-world entities and the relationships between them. Multilingual Knowledge Graph Construction (mKGC) refers to the task of automatically constructing or predicting missing entities and links for knowledge graphs in a multilingual setting. In this work, we reformulate the mKGC task as a Question Answering (QA) task and introduce mRAKL: a Retrieval-Augmented Generation (RAG) based system to perform mKGC. We achieve this by using the head entity and linking relation in a question, and having our model predict the tail entity as an answer. Our experiments focus primarily on two low-resourced languages: Tigrinya and Amharic. We experiment with using higher-resourced languages, Arabic and English, to utilize cross-lingual transfer for mKGC. With a BM25 retriever, we find that the RAG-based approach improves performance over a no-context setting. Further, our ablation studies show that with an idealized retrieval system, mRAKL improves accuracy by up to 4.92 and 8.79 percentage points for Tigrinya and Amharic, respectively.

RL-Guider: Leveraging Historical Decisions and Feedback for Drug Editing with Large Language Models

Xufeng Liu, Yixuan Ding, Jingxiang Qu, Yichi Zhang, Wenhan Gao, Yi Liu

Recent success of large language models (LLMs) in diverse domains showcases their potential to revolutionize scientific fields, including drug editing. Traditional drug editing relies on iterative conversations with domain experts, refining the drug until the desired property is achieved. This interactive and iterative process mirrors the strengths of LLMs, making them well-suited for drug editing. *In existing works, LLMs edit each molecule independently without leveraging knowledge from past edits.* However, human experts develop intuition about effective modifications over time through historical experience; accumulating past knowledge is pivotal for human experts, and so it is for LLMs. *In this work, we propose RL-Guider---a reinforcement-learning-agent to provide suggestions to LLMs; it uses the rich information provided from evaluating editing results made by the LLM based on the recommendations to improve itself over time.* RL-Guider is the first work that leverages both the comprehensive "world-level" knowledge of LLMs and the knowledge accumulated from historical feedback. As a result, RL-Guider mitigates several shortcomings of existing approaches and demonstrates superior performance.

IDEA: Enhancing the Rule Learning Ability of Large Language Model Agent through Induction, Deduction, and Abduction

Kaiyu He, Mian Zhang, Shuo Yan, Peilin Wu, Zhiyu Chen

While large language models (LLMs) have been thoroughly evaluated for deductive and inductive reasoning, their proficiency in holistic rule learning in interactive environments remains less explored. We introduce RULEARN, a novel benchmark to assess the rule-learning abilities of LLM agents in interactive settings. In RULEARN, agents strategically interact with simulated environments to gather observations, discern patterns, and solve complex problems. To enhance the rule-learning capabilities for LLM agents, we propose IDEA, a novel reasoning framework that integrates the process of **I**nduction, **De**duction, and **A**bduction. The IDEA agent generates initial hypotheses from limited observations through abduction, devises plans to validate these hypotheses or leverages them to solve problems via deduction, and refines previous hypotheses through induction, dynamically establishing and applying rules that mimic human rule-learning behaviors. Our evaluation of the IDEA framework, which involves five representative LLMs, demonstrates significant improvements over the baseline. Furthermore, our study with human participants reveals notable discrepancies in rule-learning behaviors between humans and LLMs. We believe our benchmark will serve as a valuable and challenging resource, and IDEA will provide crucial insights for the development of LLM agents capable of human-like rule learning in real-world scenarios. Our code and data have been released at: https://github.com/KaiyuHe998/RULEARN_IDEA.

ThinkGuard: Deliberative Slow Thinking Leads to Cautious Guardrails

Xiaofei Wen, Wenxuan Zhou, Wenjie Jacky Mo, Muhao Chen

Ensuring the safety of large language models (LLMs) is critical as they are deployed in real-world applications. Existing guardrails rely on rule-based filtering or single-pass classification, limiting their ability to handle nuanced safety violations. To address this, we propose ThinkGuard, a critique-augmented guardrail model that distills knowledge from high-capacity LLMs by generating structured critiques alongside safety labels. Fine-tuned on critique-augmented data, the captured deliberative thinking ability drastically enhances the guardrail's cautiousness and interpretability. Evaluated on multiple safety benchmarks, ThinkGuard achieves the highest average F1 and AUPRC, outperforming all baselines. Compared to LLaMA Guard 3, ThinkGuard improves accuracy by 16.1% and macro F1 by 27.0%. Moreover, it surpasses label-only fine-tuned models, confirming that structured critiques enhance both classification precision and nuanced safety reasoning while maintaining computational efficiency.

Dynamic Steering With Episodic Memory For Large Language Models

Van Dai Do, Quan Hung Tran, Svetha Venkatesh, Hung Le

Large Language Models (LLMs) exhibit emergent in-context learning (ICL) capabilities, allowing them to adapt to unseen tasks based on example demonstrations. Traditional ICL embeds examples within the prompt, while activation steering, uses a vector derived from examples to guide the latent states of LLMs toward desired behaviors. However, traditional ICL is difficult to control quantitatively and consumes valuable context space. Existing activation steering methods apply a single sentence-level steering vector uniformly across all tokens, ignoring LLMs' token-wise, auto-regressive nature. This coarse control can lead to inconsistencies and suboptimal adjustments during generation. To address this problem, we introduce Dynamic Steering with Episodic Memory (DSEM), a novel training-free framework that aligns LLMs to given demonstrations by steering at the token level conditioned on the input query. DSEM employs a key-value memory to store associations between generated tokens and steering vectors. During inference, it uses a nearest-neighbor mechanism to dynamically compute steering vectors for each token chunk, enabling more precise and adaptive guidance. Our method surpasses strong baselines across diverse alignment tasks - including safety, style transfer, and role-playing - demonstrating improved alignment as demonstration size scales.

Divide-Verify-Refine: Can LLMs Self-align with Complex Instructions?

Xianren Zhang, Xianfeng Tang, Hui Liu, Zongyu Wu, Qi He, Dongwon Lee, Suhang Wang

Recent studies show LLMs struggle with complex instructions involving multiple constraints (e.g., length, format, sentiment). Existing research enhances open-source LLMs using closed-source guidance (e.g., GPT-4), but this heavily relies on generated data quality. An alternative is leveraging LLMs' self-correction to refine responses for better constraint adherence. However, this is limited by the feedback quality, as we found LLMs cannot generate reliable feedback or detect errors. Moreover, the self-correction effectiveness relies on few-shot examples illustrating response modifications. As constraints in complex instructions are diverse, manually crafting such examples for each constraint type can be labor-intensive and sub-optimal. To address these two challenges, we propose the Divide-Verify-Refine (DVR) framework with three steps: (1) Divide complex instructions into single constraints and prepare appropriate tools; (2) Verify responses using tools that provide rigorous check and textual guidance (e.g., Python scripts for format checks or pre-trained classifiers for content analysis); (3) Refine: To maximize refinement effectiveness, we propose dynamic few-shot prompting, where a refinement repository collects successful refinements, and these examples are selectively retrieved for future refinements. Recognizing the lack of complexity in existing datasets, we create a new dataset of complex instructions. DVR doubles Llama3.1-8B's constraint adherence and triples Mistral-7B's performance.

Craw4LLM: Efficient Web Crawling for LLM Pretraining

Shi Yu, Zhiyuan Liu, Chenyan Xiong

Web crawl is a main source of large language models' (LLMs) pretraining data, but the majority of crawled web pages are discarded in pretraining due to low data quality. This paper presents Craw4LLM, an efficient web crawling method that explores the web graph based on the preference of LLM pretraining. Specifically, it leverages the influence of a webpage in LLM pretraining as the priority score of the web crawler's scheduler, replacing the standard graph connectivity based priority. Our experiments on a web graph containing 900 million webpages from a commercial search engine's index demonstrate the efficiency of Craw4LLM in obtaining high-quality pretraining data. With just 21% URLs crawled, LLMs pretrained on Craw4LLM data reach the same downstream performances of previous crawls, significantly reducing the crawling waste and alleviating the burdens on websites. We will make our code publicly available.

CROSSAGENTIE: Cross-Type and Cross-Task Multi-Agent LLM Collaboration for Zero-Shot Information Extraction

Meng Lu, Yuzhang Xie, Zhenyu Bi, Shuxiang Cao, Xuan Wang

Large language models (LLMs) excel in generating unstructured text. However, they struggle with producing structured output while maintaining accuracy in zero-shot information extraction (IE), such as named entity recognition (NER) and relation extraction (RE). To address these challenges, we propose CROSSAGENTIE, a multi-agent framework that enhances zero-shot IE through multi-agent LLM collaboration. CROSSAGENTIE refines LLM predictions iteratively through two mechanisms: intra-group cross-type debate, which resolves entity-label conflicts through context-based evidence and confidence aggregation, and inter-group cross-task debate, where NER and RE mutually refine outputs via bidirectional feedback. Furthermore, we introduce template fine-tuning, distilling high-confidence multi-agent outputs into a single model, significantly reducing inference cost while preserving accuracy. Experiments across five NER and five RE datasets show that CROSSAGENTIE significantly outperforms state-of-the-art zero-shot baselines by a large margin. CROSSAGENTIE effectively addresses LLMs limitations in structured prediction with an effective and efficient approach for zero-shot information extraction.

CausalAbstain: Enhancing Multilingual LLMs with Causal Reasoning for Trustworthy Abstention

Yuxi SUN, Aoqi Zuo, Wei Gao, Jing Ma

Large Language Models (LLMs) often exhibit knowledge disparities across languages. Encouraging LLMs to abstain when faced with knowledge gaps is a promising strategy to reduce hallucinations in multilingual settings. Current abstention strategies for multilingual scenarios primarily rely on generating feedback in various languages using LLMs and performing self-reflection. However, these methods can be adversely impacted by inaccuracies and biases in the generated feedback. To address this, from a causal perspective, we introduce CausalAbstain, a method that helps LLMs determine whether to utilize multiple generated feedback responses and how to identify the most useful ones. Extensive experiments demonstrate that CausalAbstain effectively selects helpful feedback and enhances abstention decisions with interpretability in both native language (Casual-native) and multilingual (Causal-multi) settings, outperforming strong baselines on two benchmark datasets covering encyclopedic and commonsense knowledge QA tasks.

EvoBench: Towards Real-world LLM-Generated Text Detection Benchmarking for Evolving Large Language Models

Xiao Yu, Yi Yu, Dongrui Liu, Kejiang Chen, Weiming Zhang, Nenghai Yu, Jing Shao

With the widespread of Large Language Models (LLMs), there has been an increasing need to detect LLM-generated texts, prompting extensive research in this area. However, existing detection methods mainly evaluate on static benchmarks, which neglect the evolving nature of LLMs. Relying on existing static benchmarks could create a misleading sense of security, overestimating the real-world effectiveness of detection methods. To bridge this gap, we introduce EvoBench, a dynamic benchmark considering a new dimension of generalization across continuously evolving LLMs. EvoBench categorizes the evolving LLMs into (1) updates over time and (2) developments like finetuning and pruning, covering 7 LLM families and their 29 evolving versions. To measure the generalization across evolving LLMs, we introduce a new EMG (Evolving Model Generalization) metric. Our evaluation of 14 detection methods on EvoBench reveals that they all struggle to maintain generalization when confronted with evolving LLMs. To mitigate the generalization problems, we further propose improvement strategies. For zero-shot detectors, we propose pruning the scoring model to extract shared features. For supervised detectors, we also propose a practical training strategy. Our research sheds light on critical challenges in real-world LLM-generated text detection and represents a significant step toward practical applications.

Benchmarking Query-Conditioned Natural Language Inference

Marc E. Canby, Xinchi Chen, Xing Niu, Jifan Chen, Bonan Min, Sergul Aydore, Vittorio Castelli

The growing excitement around the ability of large language models (LLMs) to tackle various tasks has been tempered by their propensity for generating unsubstantiated information (hallucination) and by their inability to effectively handle inconsistent inputs. To detect such issues, we propose the novel task of Query-Conditioned Natural Language Inference (QC-NLI), where the goal is to determine the semantic relationship (e.g. entailment or not entailment) between two documents conditioned on a query; we demonstrate that many common tasks regarding inconsistency detection can be formulated as QC-NLI problems. We focus on three applications in particular: fact verification, intrinsic hallucination detection, and document inconsistency detection. We convert existing datasets for these tasks into the QC-NLI format, and manual annotation confirms their high quality. Finally, we employ zero- and few-shot prompting methods to solve the QC-NLI prediction problem for each task, showing the critical importance of conditioning on the query.

Improving Efficiency in Large Language Models via Extendable Block Floating Point Representation

Dongyang Li, Zeyang Li, Bosheng Liu, Jigang Wu

Large language models (LLMs) have revolutionized natural language processing (NLP) tasks, yet their increasing size poses substantial challenges in terms of computational and memory resources. Block floating-point (BFP) arithmetic offers an effective solution by leveraging the strengths of both floating-point and fixed-point representations, leading to reductions in both storage and computational overhead. However, current low-bit BFP quantization approaches often struggle to handle extreme outliers, leading to significant accuracy degradation. To overcome this limitation, we introduce Extendable Exponent Sharing (EES), a novel BFP representation that extends the exponent bit width to capture a wider dynamic range. EES achieves this by embedding extendable exponent bits into the least significant mantissa bits, thereby increasing the shared exponent's bit width without incurring additional storage costs. To optimize the trade-off between accuracy and energy efficiency, EES employs a design space exploration strategy to optimize the configuration of extendable exponent bit widths. Experimental results show that EES outperforms representative baselines in both accuracy and computational efficiency.

DPGA-TextSyn: Differentially Private Genetic Algorithm for Synthetic Text Generation

Zhonghao Sun, Zhiliang Tian, YIPING SONG, yuyi Si, Juhua Zhang, Minlie Huang, Kai Lu, Zeyu Xiong, Xinwang Liu, Dongsheng Li

Using large language models (LLMs) has a potential risk of privacy leakage since the data with sensitive information may be used for fine-tuning the LLMs. Differential privacy (DP) provides theoretical guarantees of privacy protection, but its practical application in LLMs still has the problem of privacy-utility trade-off. Researchers synthesized data with strong generation capabilities closed-source LLMs (i.e., GPT-4) under DP to alleviate this problem, but this method is not so flexible in fitting the given privacy distributions without fine-tuning. Besides, such methods can hardly balance the diversity of synthetic data and its relevance to target privacy data without accessing so much private data. To this end, this paper proposes DPGA-TextSyn, combining general LLMs with genetic algorithm (GA) to produce relevant and diverse synthetic text under DP constraints. First, we integrate the privacy gene (i.e., metadata) to generate better initial samples. Then, to achieve survival of the fittest and avoid homogeneity, we use privacy nearest neighbor voting and similarity suppression to select elite samples. In addition, we expand elite samples via genetic strategies such as mutation, crossover, and generation to expand the search scope of GA. Experiments show that this method significantly improves the performance of the model in downstream tasks while ensuring privacy.

mStyleDistance: Multilingual Style Embeddings and their Evaluation

Justin Qiu, Jiacheng Zhu, Ajay Patel, Marianna Apidianaki, Chris Callison-Burch

Style embeddings are useful for stylistic analysis and style transfer, yet they only exist for English. We introduce Multilingual StyleDistance (mStyleDistance), a method that can generate style embeddings in new languages using synthetic data and a contrastive loss. We create style embeddings in nine languages and a multilingual STEL-or-Content benchmark (Wegmann et al., 2022) that serves to assess their quality. We also employ our embeddings in an authorship verification task involving different languages. Our results show that mStyleDistance embeddings outperform existing style embeddings on these benchmarks and generalize well to unseen features and languages. We make our models and datasets publicly available.

MAGIC-VQA: Multimodal And Grounded Inference with Commonsense Knowledge for Visual Question Answering

Shuo Yang, Siwen Luo, Caren Han, Eduard Hovy

Visual Question Answering (VQA) necessitates models to reason effectively across visual and textual modalities. However, existing Large Vision-Language Models (LVLMs) often fall short in achieving human-like reasoning due to a lack of integrated commonsense knowledge, limiting their robustness and accuracy in real-world scenarios where both explicit facts and implicit understanding are crucial. To address this challenge, we present MAGIC-VQA: Multimodal And Grounded Inference with Commonsense Knowledge, a novel framework designed to enhance multimodal inference by integrating commonsense reasoning. MAGIC-VQA introduces a three-stage process: (1) Explicit Commonsense Knowledge Retrieval from external knowledge graphs, (2) By-Type Commonsense Knowledge Post-Processing to refine contextual relevance, and (3) Implicit Commonsense Knowledge Augmentation using a heterogeneous graph processed by a Graph Neural Network (GNN). These stages collectively enable nuanced, context-aware reasoning without extensive pre-training or intricate prompt tuning. Our MAGIC-VQA significantly improves comprehensive benchmark datasets, surpassing existing models in tasks requiring advanced commonsense reasoning. MAGIC-VQA establishes a robust pathway for integrating commonsense knowledge into VQA, bridging the gap between vision-language inputs and high-level reasoning for improved reliability and contextual accuracy.

Natural Logic at the Core: Dynamic Rewards for Entailment Tree Generation

Jihao Shi, Xiao Ding, Kai Xiong, Hengwei Zhao, Bing Qin, Ting Liu

Entailment trees are essential for enhancing interpretability and transparency in tasks like question answering and natural language understanding. However, existing approaches often lack logical consistency, as they rely on static reward structures or ignore the intricate dependencies within multi-step reasoning. To address these limitations, we propose a method that integrates natural logic principles into reinforcement learning, enabling dynamic reward computation to guide entailment tree generation. Our approach ensures logical consistency across reasoning steps while improving interpretability and generalization. Experiments on EntailmentBank demonstrate significant improvements over state-of-the-art methods, highlighting the effectiveness of natural logic in structured reasoning.

Each graph is a new language: Graph Learning with LLMs

Huachi Zhou, Jiahe Du, Chuang Zhou, Chang Yang, Yilin Xiao, Yuxuan Xie, Xiao Huang

Natural language has been extensively used for modeling text-attributed graphs with LLMs. Natural language is used to describe the graph for LLMs to understand or serve as component of the graph, e.g., textual attributes for embedding generation. However, natural language is inherently redundant and unstructured, making it unsuitable for modeling high-order neighbors with LLMs. Specifically, (i) graph descriptions become verbose, overwhelming LLMs, and (ii) only relying on attribute embeddings limits LLM's ability to capture the adequate graph structural information. These limitations make it difficult to model graphs both concisely and adequately using sole natural language with LLMs. Inspired by the observation that LLMs pre-trained on one language can achieve exceptional performance on another with minimal additional training, we propose Graph-Defined Language for Large Language Model (GDL4LLM). This novel framework enables LLMs to transfer their powerful language understanding capabilities to graph-structured data. GDL4LLM translates the graph into a graph language corpus instead of graph descriptions and pre-trains LLMs on this corpus to adequately understand the graph. This corpus represents the subgraph centered around target nodes concisely with only a few tokens during fine-tuning on downstream tasks. By treating the graph as a new language, GDL4LLM enables LLMs to model text-attributed graph adequately and concisely. Extensive experiments on five datasets demonstrate that GDL4LLM outperforms description-based and embedding-based baselines by efficiently modeling different orders of neighbors.

ShieldHead: Decoding-time Safeguard for Large Language Models

Zitao Xuan, Xiaofeng Mao, Da Chen, Xin Zhang, Yuhan Dong, JUN ZHOU

In light of the widespread deployment of Large Language Models (LLMs), the responsibility for safeguarding and regulating LLM-generated content has taken on heightened significance. Recent advancements in LLM-based moderation methods, e.g., LlamaGuard, have demonstrated remarkable promise in identifying safety risks associated with both inputs and outputs in human-AI interactions. However, integrating LLM-based safeguards into a chatbot system requires an additional inference stage involving a moderation LLM with billions of parameters, which significantly increases computational costs and reduces overall efficiency. In this paper, we demonstrate that simply learning a classification head on the last-layer hidden states of the dialogue model provides a strong capability to identify harmful contents. The classification head, referred to as ShieldHead, serves as an auxiliary branch paralleled with next-token-prediction LM head, enabling the detection of potential risks in past text sequences. Additionally, a label disambiguation technique is employed to supervise ShieldHead with both token-level and sentence-level labels, which further enhances its performance. ShieldHead exhibits remarkable efficiency during inference, providing real-time moderation results alongside token-wise streaming output during the chatbot system's decoding phase. Extensive experimental results demonstrate the superiority of the proposed framework: a state-of-the-art performance on the XSTest and SafeRLHF datasets while running at a speed about **300×** faster (**<1ms**) than previous LLM-based moderation models with **99%** less parameters of LlamaGuard.

Same Company, Same Signal: The Role of Identity in Earnings Call Transcripts

Ding Yu, Zhuo Liu, Hangfeng He

Post-earnings volatility prediction is critical for investors, with previous works often leveraging earnings call transcripts under the assumption that their rich semantics contribute significantly. To further investigate how transcripts impact volatility, we introduce DEC, a dataset featuring accurate volatility calculations enabled by the previously overlooked beforeAfterMarket attribute and dense ticker coverage. Unlike established benchmarks, where each ticker has only around two earnings, DEC provides 20 earnings records per ticker. Using DEC, we reveal that post-earnings volatility undergoes significant shifts, with each ticker displaying a distinct volatility distribution. To leverage historical post-earnings volatility and capture ticker-specific patterns, we propose two training-free baselines: Post-earnings Volatility (PEV) and Same-ticker Post-earnings Volatility (STPEV). These baselines surpass all transcripts-based models on DEC as well as on established benchmarks. Additionally, we demonstrate that current transcript representations predominantly capture ticker identity rather than offering financially meaningful insights specific to each earnings. This is evidenced by two key observations: earnings representations from the same ticker exhibit significantly higher similarity compared to those from different tickers, and predictions from transcript-based models show strong correlations with prior post-earnings volatility.

Understanding and Meeting Practitioner Needs When Measuring Representational Harms Caused by LLM-Based Systems

Emma Harvey, Emily Sheng, Su Lin Blodgett, Alexandra Chouldechova, Jean Garcia-Gathright, Alexandra Olteanu, Hanna Wallach

The NLP research community has made publicly available numerous instruments for measuring representational harms caused by large language model (LLM)-based systems. These have taken the form of datasets, metrics, and other tools designed to measure harms such as stereotyping and demeaning language. In this study we examine the extent to which such instruments meet the needs of practitioners tasked with evaluating LLM-based systems. Via semi-structured interviews with AI practitioners (N=12), we find that practitioners are often unable to use publicly available instruments for measuring representational harms. We identify two types of challenges. In some cases, instruments are not useful because they do not meaningfully measure what practitioners seek to measure or are otherwise misaligned with practitioner needs. In other cases, instruments - even useful instruments - are not used by practitioners due to institutional barriers impeding their uptake. Drawing on measurement theory and pragmatic measurement, we provide recommendations for addressing these challenges to better meet practitioner needs.

Biases Propagate in Encoder-based Vision-Language Models: A Systematic Analysis From Intrinsic Measures to Zero-shot Retrieval Outcomes

Kshitish Ghate, Tessa Charlesworth, Mona T. Diab, Aylin Caliskan

Understanding how social-group biases intrinsic to foundational encoder-based vision-language models (VLMs) manifest in downstream tasks has significant implications for fair AI systems. In this study, we demonstrate that intrinsic biases in VLM representations systematically ``carry over'' or propagate into zero-shot retrieval tasks, revealing how deeply rooted biases shape a model's outputs. We introduce a controlled framework to measure this propagation by correlating (a) intrinsic measures of bias in the representational space with (b) extrinsic measures of bias in zero-shot text-to-image (TTI) and image-to-text (ITT) retrieval. Results show substantial correlations between intrinsic and extrinsic bias, with an average \rho = 0.83 \pm 0.10. This pattern is consistent across 114 analyses, both retrieval directions, six social groups, and three distinct VLMs. Notably, we find that larger/better-performing models exhibit greater bias propagation, a finding that raises concerns given the trend towards increasingly complex AI models. Our framework introduces baseline evaluation tasks to measure the propagation of group and valence signals. Investigations reveal that underrepresented groups experience less robust propagation, further skewing their model-related outcomes.

Let The Jury Decide: Fair Demonstration Selection for In-Context Learning through Incremental Greedy Evaluation

Sadaf MD Halim, Chen Zhao, Xintao Wu, Latifur Khan, Christan Grant, Fariha Ishrat Rahman, Feng Chen

Large Language Models (LLMs) are powerful in-context learners, achieving strong performance with just a few high-quality demonstrations. However, fairness concerns arise in many in-context classification tasks, especially when predictions involve sensitive attributes. To address this, we propose JUDGE—a simple yet effective framework for selecting fair and representative demonstrations that improve group fairness in In-Context Learning. JUDGE constructs the demonstration set iteratively using a greedy approach, guided by a small, carefully selected jury set. Our method remains robust across varying LLM architectures and datasets, ensuring consistent fairness improvements. We evaluate JUDGE on four datasets using four LLMs, comparing it against seven baselines. Results show that JUDGE consistently improves fairness metrics without compromising accuracy.

Speech Act Patterns for Improving Generalizability of Explainable Politeness Detection Models

Ahmad Aljanaideh

The lack of explainability in state-of-the-art Natural Language Understanding (NLU) classification models has increased interest in developing techniques for improving explainable linear feature-based models (e.g., Logistic Regression/SVM). Politeness detection is a task that exemplifies this interest. While those techniques perform well on the task when applied to data from the same domain as the training data, they lack generalizability and thus fall short when applied to data from other domains. This is due to their reliance on discovering domain-specific word-level features. We introduce a method for improving the generalizability of explainable politeness models by relying on speech act patterns instead of words, leveraging speech act labels assigned by the GPT-4 model. This approach goes beyond the mere words and injects intent into politeness classification models, enhancing their generalizability. Results demonstrate that the proposed method achieves state-of-the-art accuracy in the cross-domain setting among explainable methods, while falling short in the in-domain setting. Our findings illustrate that explainable models can benefit from Large Language Models.

V-ALPHASOCIAL: Benchmark and Self-Reflective Chain-of-Thought Generation for Visual Social Commonsense Reasoning

Zongyu Lin, Zhikun Xu, Xiaohan Song, Yixin Wan, Xingcheng Yao, Tsung-Han Lin, Selina Song, Pranav Subbaraman, Ben Zhou, Kai-Wei Chang, Yizhou Sun

Social commonsense reasoning naturally involves both the verbal and non-verbal cues of a social interaction. It is important for Large Vision-Language Models (VLMs) to leverage both textual and visual information in performing tasks like social understanding and reasoning. However, while current LLMs have shown good social reasoning capabilities in textual context, whether they can effectively incorporate visual information in social comprehension remains under-explored. To narrow the gap, we first construct and propose a benchmark: V-Social, featuring well-aligned text and visual content, tailored to assess visual social commonsense for multimodal foundation models. Through experimenting with V-Social, we find that even the most advanced VLM, GPT-4o, often falls short in social commonsense reasoning. This highlights the critical need to enhance the social grounding of VLMs. One major obstacle for improving this is the lack of high-quality data with good reasoning process. To overcome this obstacle, we introduce V-AlphaSocial, a novel method that generates high-quality chain-of-thought reasoning paths from unlabeled data. We design a visual reasoning reward model to improve VLM, and then iteratively refine both the VLM and the reward model. Our extensive analysis showcases how our method enhances social commonsense reasoning, proposing an effective approach that facilitates deeper exploration into field.

Outlier-weighed Layerwise Sampling for LLM Fine-tuning

Pengxiang Li, Lu Yin, Xiaowei Gao, Shiwei Liu

The rapid advancements in Large Language Models (LLMs) have revolutionized various natural language processing tasks. However, the substantial size of LLMs presents significant challenges in training or fine-tuning. While parameter-efficient approaches such as low-rank adaptation (LoRA) have gained popularity, they often compromise performance compared to full-rank fine-tuning. In this paper, we propose Outlier-weighed Layerwise Sampling (OWS), a new memory-efficient fine-tuning approach, inspired by the layerwise outlier distribution of LLMs. Unlike LoRA, which adds extra adapters to all layers, OWS strategically assigns higher sampling probabilities to layers with more outliers, selectively sampling only a few layers and fine-tuning their pre-trained weights. To further increase the number of fine-tuned layers without a proportional rise in memory costs, we incorporate gradient low-rank projection, further boosting the approach's performance. Our extensive experiments across various architectures, including LLaMa2 and Mistral, demonstrate that OWS consistently outperforms baseline approaches, including full fine-tuning. Specifically, it achieves up to a 1.1% average accuracy gain on the Commonsense Reasoning benchmark, a 3.0% improvement on MMLU, and a notable 10% boost on MT-Bench, while being more memory efficient. OWS allows us to fine-tune 7B LLMs with only 21GB of memory. Our code is submitted.

Core: Robust Factual Precision with Informative Sub-Claim Identification

Zhengping Jiang, Jingyu Zhang, Nathaniel Weir, Seth Ebner, Miriam Wanner, Kate Sanders, Daniel Khashabi, Anqi Liu, Benjamin Van Durme

Hallucinations pose a challenge to the application of large language models (LLMs) thereby motivating the development of metrics to evaluate factual precision. We observe that popular metrics using the Decompose-Then-Verify framework, such as FActScore, can be manipulated by adding obvious or repetitive subclaims to artificially inflate scores. This observation motivates our new customizable plug-and-play subclaim selection component called Core, which filters down individual subclaims according to their uniqueness and informativeness. We show that many popular factual precision metrics augmented by Core are substantially more robust on a wide range of knowledge domains. We release an evaluation framework supporting easy and modular use of Core and various decomposition strategies, which we recommend adoption by the community. We also release an expansion of the FActScore biography dataset to facilitate further studies of decomposition-based factual precision evaluation.

DYNTEXT: Semantic-Aware Dynamic Text Sanitization for Privacy-Preserving LLM Inference

Juhua Zhang, Zhiliang Tian, Minghang Zhu, YIPING SONG, Minlie Huang, Siyi Yang, Qiunan Du, Xinwang Liu, Taishu sheng, Dongsheng Li

LLMs face privacy risks when handling sensitive data. To ensure privacy, researchers use differential privacy (DP) to provide protection by adding noise during LLM training. However, users may be hesitant to share complete data with LLMs. Researchers follow local DP to sanitize the text on the user side and feed non-sensitive text to LLMs. The sanitization usually uses a fixed non-sensitive token list or a fixed noise distribution, which induces the risk of being attacked or semantic distortion. We argue that the token's protection level should be adaptively adjusted according to its semantic-based information to balance the privacy-utility trade-off. In this paper, we propose DYNTEXT, an LDP-based Dynamic Text sanitization for privacy-preserving LLM inference, which dynamically constructs semantic-aware adjacency lists of sensitive tokens to sample non-sensitive tokens for perturbation. Specifically, DYNTEXT first develops a semantic-based density modeling under DP to extract each token's density information. We propose token-level smoothing sensitivity by combining the idea of global sensitivity (GS) and local sensitivity (LS), which dynamically adjusts the noise scale to avoid excessive noise in GS and privacy leakage in LS. Then, we dynamically construct an adjacency list for each sensitive token based on its semantic density information. Finally, we apply the replacement mechanism to sample non-sensitive, semantically similar tokens from the adjacency list to replace sensitive tokens. Experiments show that DYNTEXT excels strong baselines on three datasets.

Localizing and Mitigating Errors in Long-form Question Answering

Rachneet Singh Sachdeva, Yixiao Song, Mohit Iyyer, Iryna Gurevych

Long-form question answering (LFQA) aims to provide thorough and in-depth answers to complex questions, enhancing comprehension. However, such detailed responses are prone to hallucinations and factual inconsistencies, challenging their faithful evaluation. This work introduces HaluQuestQA, the first hallucination dataset with localized error annotations for human-written and model-generated LFQA answers. HaluQuestQA comprises 698 QA pairs with 1.8k span-level error annotations for five different error types by expert annotators, along with preference judgments. Using our collected data, we thoroughly analyze the shortcomings of long-form answers and find that they lack comprehensiveness and provide unhelpful references. We train an automatic feedback model on this dataset that predicts error spans with incomplete information and provides associated explanations. Finally, we propose a prompt-based approach, Error-Informed Refinement, that uses signals from the learned feedback model to refine generated answers, which we show reduces errors and improves the quality of the answers across multiple models. Furthermore, humans find the answers generated by our approach comprehensive and highly prefer them (84%) over the baseline answers.

TRATES: Trait-Specific Rubric-Assisted Cross-Prompt Essay Scoring

Sohaila Eltanbouly, Salam Albatarni, Tamer Elsayed

Research on holistic Automated Essay Scoring (AES) is long-dated; yet, there is a notable lack of attention for assessing essays according to individual traits. In this work, we propose TRATES, a novel trait-specific and rubric-based cross-prompt AES framework that is generic yet specific to the underlying trait. The framework leverages a Large Language Model (LLM) that utilizes the trait grading rubrics to generate trait-specific features (represented by assessment questions), then assesses those features given an essay. The trait-specific features are eventually combined with generic writing-quality and prompt-specific features to train a simple classical regression model that predicts trait scores of essays from an unseen prompt. Experiments show that TRATES achieves a new state-of-the-art performance across all traits on a widely-used dataset, with the generated LLM-based features being the most significant.

DiffLM: Controllable Synthetic Data Generation via Diffusion Language Models

Ying Zhou, Xinyao Wang, Yulei Niu, Yaojie Shen, Lexin Tang, Fan Chen, Ben He, Le Sun, Longyin Wen

Recent advancements in large language models (LLMs) have significantly enhanced their knowledge and generative capabilities, leading to a surge of interest in leveraging LLMs for high-quality data synthesis. However, synthetic data generation via prompting LLMs remains challenging due to LLMs' limited understanding of target data distributions and the complexity of prompt engineering, especially for structured formatted data. To address these issues, we introduce DiffLM, a controllable data synthesis framework based on variational autoencoder (VAE), which further (1) leverages diffusion models to reserve more information of original distribution and format structure in the learned latent distribution and (2) decouples the learning of target distribution knowledge from the LLM's generative objectives via a plug-and-play latent feature injection module. As we observed significant discrepancies between the VAE's latent representations and the real data distribution, the latent diffusion module is introduced into our framework to learn a fully expressive latent distribution. Evaluations on seven real-world datasets with structured formatted data (i.e., Tabular, Code and Tool data) demonstrate that DiffLM generates high-quality data, with performance on downstream tasks surpassing that of real data by 2%–7% in certain cases. Data and code will be released upon acceptance.

Exploring Jailbreak Attacks on LLMs through Intent Concealment and Diversion

Tiehan Cui, Yanxu Mao, Peipei Liu, Congying Liu, Datao You

Although large language models (LLMs) have achieved remarkable advancements, their security remains a pressing concern. One major threat is jailbreak attacks, where adversarial prompts bypass model safeguards to generate harmful or objectionable content. Researchers study jailbreak attacks to understand security and robustness of LLMs. However, existing jailbreak attack methods face two main challenges: (1) an excessive number of iterative queries, and (2) poor generalization across models. In addition, recent jailbreak evaluation datasets focus primarily on question-answering scenarios, lacking attention to text generation tasks that require accurate regeneration of toxic content. To tackle these challenges, we propose two contributions: (1) **ICE**, a novel black-box jailbreak method that employs **I**ntent **C**oncealment and div**E**rsion to effectively circumvent security constraints. **ICE** achieves high attack success rates (ASR) with a single query, significantly improving efficiency and transferability across different models. (2) **BiSceneEval**, a comprehensive dataset designed for assessing LLM robustness in question-answering and text-generation tasks. Experimental results demonstrate that **ICE** outperforms existing jailbreak techniques, revealing critical vulnerabilities in current defense mechanisms. Our findings underscore the necessity of a hybrid security strategy that integrates predefined security mechanisms with real-time semantic decomposition to enhance the security of LLMs.

Bandit-Based Prompt Design Strategy Selection Improves Prompt Optimizers

Rin Ashizawa, Yoichi Hirose, Nozomu Yoshinari, Kento Uchida, Shinichi Shirakawa

Prompt optimization aims to search for effective prompts that enhance the performance of large language models (LLMs). Although existing prompt optimization methods have discovered effective prompts, they often differ from sophisticated prompts carefully designed by human experts. Prompt design strategies, representing best practices for improving prompt performance, can be key to improving prompt optimization. Recently, a method termed the Autonomous Prompt Engineering Toolbox (APET) has incorporated various prompt design strategies into the prompt optimization process. In APET, the LLM is needed to implicitly select and apply the appropriate strategies because prompt design strategies can have negative effects. This implicit selection may be suboptimal due to the limited optimization capabilities of LLMs. This paper introduces Optimizing Prompts with sTrategy Selection (OPTS), which implements explicit selection mechanisms for prompt design. We propose three mechanisms, including a Thompson sampling-based approach, and integrate them into EvoPrompt, a well-known prompt optimizer. Experiments optimizing prompts for two LLMs, Llama-3-8B-Instruct and GPT-4o mini, were conducted using BIG-Bench Hard. Our results show that the selection of prompt design strategies improves the performance of EvoPrompt, and the Thompson sampling-based mechanism achieves the best overall results.

Is Large Language Model Performance on Reasoning Tasks Impacted by Different Ways Questions Are Asked?

Seok Hwan Song, Mohna Chakraborty, Qi Li, Wallapak Tavanapong

Large Language Models (LLMs) have been evaluated using diverse question types, e.g., multiple-choice, true/false, and short/long answers. This study answers an unexplored question about the impact of different question types on LLM accuracy on reasoning tasks. We investigate the performance of five LLMs on three different types of questions using quantitative and deductive reasoning tasks. The performance metrics include accuracy in the reasoning steps and choosing the final answer. Key Findings: (1) Significant differences exist in LLM performance across different question types. (2) Reasoning accuracy does not necessarily correlate with the final selection accuracy. (3) The number of options and the choice of words, influence LLM performance.

MutantPrompt: Prompt Optimization via Mutation Under a Budget on Modest-sized LMs

Arijit Nag, Animesh Mukherjee, Niloy Ganguly, Soumen Chakrabarti

Prompts serve as a critical instruction interface to unlock the diverse capabilities of Large Language Models (LLMs), thus directly influencing the quality of their outputs. While prompt engineering has shown great promise, identifying optimal prompts remains a significant challenge, particularly for low-resource languages, which often face higher computational costs due to increased token generation and limited gold standard task data. In response, we propose MutantPrompt, a framework that leverages multi-armed bandit algorithms to efficiently identify optimal prompts tailored to low-resource languages. By framing prompt selection as an exploration-exploitation problem under a fixed computational budget, the framework dynamically balances exploring new prompts with exploiting known high-performing ones. We demonstrate the framework's effectiveness across multiple low-resource Indic language tasks, including classification, question-answering and causal reasoning using three small parameter-size LLMs. The results highlight the cost efficiency of the search method in finding optimal prompts and resulting performance improvements.

Heuristic-based Search Algorithm in Automatic Instruction-focused Prompt Optimization: A Survey

Wendi Cui, Jiaxin Zhang, Zhuohang Li, Hao Sun, Damien Lopez, Kamalika Das, Bradley A. Malin, Sricharan Kumar

Recent advances in Large Language Models(LLMs) have led to remarkable achievements across a variety of Natural Language Processing(NLP) tasks, making prompt engineering increasingly central to guiding model outputs. While manual methods (e.g., "chain-of-thought," "step-by-step" prompts) can be effective, they typically rely on intuition and do not automatically refine prompts over time. In contrast, automatic prompt optimization employing heuristic-based search algorithms can systematically explore and improve prompts with minimal human oversight. This survey proposes a comprehensive taxonomy of these methods, categorizing them by where optimization occurs, what is optimized, what criteria drive the optimization, which operators generate new prompts, and which iterative search algorithms are applied. We further highlight specialized datasets and tools that support and accelerate automated prompt refinement. We conclude by discussing key open challenges, pointing toward future opportunities for more robust and versatile LLM applications.

CONSENSAGENT: Towards Efficient and Effective Consensus in Multi-Agent LLM Interactions Through Sycophancy Mitigation

Priya Pitre, Naren Ramakrishnan, Xuan Wang

Multi-agent large language model (LLM) systems have shown remarkable performance in tasks such as reasoning, planning, and decision-making. However, their applicability is limited by challenges such as high computational costs and robustness issues. In this work, we identify and systematically evaluate a critical yet overlooked challenge: sycophancy, where agents reinforce each other's responses instead of critically engaging with the debate. This behavior inflates computational costs by requiring additional debate rounds to reach consensus, limiting the efficiency of multi-agent LLM systems. Through experiments on six benchmark reasoning datasets across three models, we analyze the impact of sycophancy and its role in reducing the reliability of multi-agent debate. Motivated by our findings, we propose CONSENSAGENT, a novel framework that dynamically refines prompts based on agent interactions to mitigate sycophancy. CONSENSAGENT reduces the number of debate rounds required for consensus and improves accuracy while maintaining efficiency. It significantly outperforms both single-agent and multi-agent baselines, achieving state-of-the-art results across all benchmark datasets. Our findings highlight the crucial role of structured prompt optimization in multi-agent setups and establish a foundation for more reliable, efficient multi-agent LLM systems in real-world applications.

CausalLink: An Interactive Evaluation Framework for Causal Reasoning

Jinyue Feng, Frank Rudzicz

We present CausalLink, an innovative evaluation framework that interactively assesses the causal reasoning skill to identify the correct intervention in conversational language models. Each CausalLink test case creates a hypothetical environment in which the language models are instructed to apply interventions to entities whose interactions follow predefined causal relations generated from controllable causal graphs. Our evaluation framework isolates causal capabilities from the confounding effects of world knowledge and semantic cues. We evaluate a series of LLMs in a scenario featuring movements of geometric shapes and discover that models start to exhibit reliable reasoning on two or three variables at the 14-billion-parameter scale. However, the performance of state-of-the-art models such as GPT4o degrades below random chance as the number of variables increases. We identify and analyze several key failure modes.

Creating a Lens of Chinese Culture: A Multimodal Dataset for Chinese Pun Rebus Art Understanding

Tuo Zhang, Tiantian Feng, Yibin Ni, Mengqin Cao, Ruying Liu, Kiana Avestimehr, Katharine Butler, Yanjun Weng, Mi Zhang, Shrikanth Narayanan, Salman Avestimehr

Large vision-language models (VLMs) have demonstrated remarkable abilities in understanding everyday content. However, their performance in the domain of art, particularly culturally rich art forms, remains less explored. As a pearl of human wisdom and creativity, art encapsulates complex cultural narratives and symbolism. In this paper, we offer the Pun Rebus Art Dataset, a multimodal dataset for art understanding deeply rooted in traditional Chinese culture. We focus on three primary tasks: identifying salient visual elements, matching elements with their symbolic meanings, and explanations for the conveyed messages. Our evaluation reveals that state-of-the-art VLMs struggle with these tasks, often providing biased and hallucinated explanations and showing limited improvement through in-context learning. By releasing the Pun Rebus Art Dataset, we aim to facilitate the development of VLMs that can better understand and interpret culturally specific content, promoting greater inclusiveness beyond English-based corpora. The dataset and evaluation code are available at [this link](https://github.com/zhang-tuo-pdf/Pun-Rebus-Art-Benchmark).

Building A Proof-Oriented Programmer That Is 64% Better Than GPT-4o Under Data Scarcity

Dylan Zhang, Justin Wang, Tianran Sun

Existing LMs struggle with proof-oriented programming due to data scarcity, which manifest in two key ways: (1) a lack of sufficient corpora for proof-oriented programming languages such as F*, and (2) the absence of large-scale, project-level proof-oriented implementations that can teach the model the intricate reasoning process when performing proof-oriented programming. We present the first on synthetic data augmentation for project level proof oriented programming for both generation and repair. Our method addresses data scarcity by synthesizing basic proof-oriented programming problems for proficiency in that language; incorporating diverse coding data for reasoning capability elicitation and creating new proofs and repair data within existing repositories. This approach enables language models to both synthesize and repair proofs for function- and repository-level code. We show that our fine-tuned 14B parameter model, PoPilot, can exceed the performance of the models that outperforms GPT-4o in project-level proof-oriented programming by 64% relative margin, and can improve GPT-4o's performance by 54% by repairing its outputs over GPT-4o's self-repair.

Diversification Catalyzes Language Models' Instruction Generalization To Unseen Semantics

Dylan Zhang, Justin Wang, Francois Charton

Instruction-tuned language models excel in knowledge, reasoning, and instruction-following. While knowledge and reasoning are well-explored, the factors enabling generalization to unseen instructions remain underexplored due to challenges in isolating instruction-following dynamics. In this work, we model instruction-following as a computational process and design controlled experiments inspired by the Turing-complete Markov algorithm to disentangle its dynamics. Our findings reveal that the ability to generalize to instructions with unseen semantics emerges only when training data is strategically diversified across rich semantics. This finding gives us the hammer that breaks down the wall separating training instructions from unseen ones encountered in the wild. For specialist models, a balanced mix of in-domain and diverse out-of-domain tasks enhances performance more effectively than simply increasing in-domain data. For generalist models, domain diversification consistently outweighs the costs of reduced task-specific data, regardless of data budgets. Furthermore, we show that proper diversification with a lower data budget can outperform simply scaling up data volume. These findings highlight strategic data diversification as key to optimizing instruction-following and improving model performance across applications.

Revisiting 3D LLM Benchmarks: Are We Really Testing 3D Capabilities?

Jiahe Jin, Yanheng He, Mingyan Yang

In this work, we identify the "2D-Cheating" problem in 3D LLM evaluation, where these tasks might be easily solved by VLMs with rendered images of point clouds, exposing ineffective evaluation of 3D LLMs' unique 3D capabilities. We test VLM performance across multiple 3D LLM benchmarks and, using this as a reference, propose principles for better assessing genuine 3D understanding. We also advocate explicitly separating 3D abilities from 1D or 2D aspects when evaluating 3D LLMs.

Safeguarding RAG Pipelines with GMTP: A Gradient-based Masked Token Probability Method for Poisoned Document Detection

San Kim, Jonghwi Kim, Yejin Jeon, Gary Lee

Retrieval-Augmented Generation (RAG) enhances Large Language Models (LLMs) by providing external knowledge for accurate and up-to-date responses. However, this reliance on external sources exposes a security risk; attackers can inject poisoned documents into the knowledge base to steer the generation process toward harmful or misleading outputs. In this paper, we propose Gradient-based Masked Token Probability (GMTP), a novel defense method to detect and filter out adversarially crafted documents. Specifically, GMTP identifies high-impact tokens by examining gradients of the retriever's similarity function. These key tokens are then masked, and their probabilities are checked via a Masked Language Model (MLM). Since injected tokens typically exhibit markedly low masked-token probabilities, this enables GMTP to easily detect malicious documents and achieve high-precision filtering. Experiments demonstrate that GMTP is able to eliminate over 90% of poisoned content while retaining relevant documents, thus maintaining robust retrieval and generation performance across diverse datasets and adversarial settings.

HATA: Trainable and Hardware-Efficient Hash-Aware Top-k Attention for Scalable Large Model Inference

Ping Gong, Jiawei Yi, Shengnan Wang, Juncheng Zhang, Zewen Jin, Ouxiang Zhou, Ruibo Liu, Guanbin Xu, Youhui Bai, Bowen Ye, Kun Yuan, Tong Yang, Gong Zhang, Renhai Chen, Feng Wu, Cheng Li

Large Language Models (LLMs) have emerged as a pivotal research area, yet the attention module remains a critical bottleneck in LLM inference, even with techniques like KVCache to mitigate redundant computations. While various top-k attention mechanisms have been proposed to accelerate LLM inference by exploiting the inherent sparsity of attention, they often struggled to strike a balance between efficiency and accuracy. In this paper, we introduce HATA (Hash-Aware Top-k Attention), a novel approach that systematically integrates low-overhead learning-to-hash techniques into the Top-k attention process. Different from the existing top-k attention methods which are devoted to seeking an absolute estimation of qk score, typically with a great cost, HATA maps queries and keys into binary hash codes, and acquires the relative qk score order with a quite low cost, which is sufficient for realizing top-k attention. Extensive experiments demonstrate that HATA achieves up to 7.2\times speedup compared to vanilla full attention while maintaining model accuracy. In addition, HATA outperforms the state-of-the-art top-k attention methods in both accuracy and efficiency across multiple mainstream LLM models and diverse tasks. To foster academic collaboration, we will open-source the HATA implementation soon.

HPSS: Heuristic Prompting Strategy Search for LLM Evaluators

Bosi Wen, Pei Ke, Yufei Sun, Cunxiang Wang, Xiaotao Gu, Jinfeng Zhou, Jie Tang, Hongning Wang, Minlie Huang

Since the adoption of large language models (LLMs) for text evaluation has become increasingly prevalent in the field of natural language processing (NLP), a series of existing works attempt to optimize the prompts for LLM evaluators to improve their alignment with human judgment. However, their efforts are limited to optimizing individual factors of evaluation prompts, such as evaluation criteria or output formats, neglecting the combinatorial impact of multiple factors, which leads to insufficient optimization of the evaluation pipeline. Nevertheless, identifying well-behaved prompting strategies for adjusting multiple factors requires extensive enumeration. To this end, we comprehensively integrate 8 key factors for evaluation prompts and propose a novel automatic prompting strategy optimization method called Heuristic Prompting Strategy Search (HPSS). Inspired by the genetic algorithm, HPSS conducts an iterative search to find well-behaved prompting strategies for LLM evaluators. A heuristic function is employed to guide the search process, enhancing the performance of our algorithm. Extensive experiments across four evaluation tasks demonstrate the effectiveness of HPSS, consistently outperforming both human-designed evaluation prompts and existing automatic prompt optimization methods.

LegalCore: A Dataset for Event Coreference Resolution in Legal Documents

Kangda Wei, Xi Shi, Jonathan Tong, Sai Ramana Reddy, Anandhavelu Natarajan, Rajiv Jain, Aparna Garimella, Ruihong Huang

Recognizing events and their coreferential mentions in a document is essential for understanding semantic meanings of text. The existing research on event coreference resolution is mostly limited to news articles. In this paper, we present the first dataset for the legal domain, LegalCore, which has been annotated with comprehensive event and event coreference information. The legal contract documents we annotated in this dataset are several times longer than news articles, with an average length of around 25k tokens per document. The annotations show that legal documents have dense event mentions and feature both short-distance and super long-distance coreference links between event mentions. We further benchmark mainstream Large Language Models (LLMs) on this dataset for both event detection and event coreference resolution tasks, and find that this dataset poses significant challenges for state-of-the-art open-source and proprietary LLMs, which perform significantly worse than a supervised baseline. We will publish the dataset as well as the code.

MemeDetoxNet: Balancing Toxicity Reduction and Context Preservation

Gitanjali Kumari, Jitendra solanki, Asif Ekbal

Toxic memes often spread harmful and offensive content and pose a significant challenge in online environments. In this paper, we present MemeDetoxNet, a robust framework designed to mitigate toxicity in memes by leveraging fine-tuned pre-trained models. Our approach utilizes the interpretability of CLIP (Contrastive Language-Image Pre-Training) to identify toxic elements within the visual and textual components of memes. Our objective is to automatically assess the immorality of toxic memes and transform them into morally acceptable alternatives by employing large language models (LLMs) to replace offensive text and blurring toxic regions in the image. As a result, we proposed MemeDetoxNet that has three main primitives: (1) detection of toxic memes, (2) localizing and highlighting toxic visual and textual attributes, and (3) manipulating the toxic content to create a morally acceptable alternative. Empirical evaluation on several publicly available meme datasets shows a reduction in toxicity by approximately 10-20%. Both qualitative and quantitative analyses further demonstrate MemeDetoxNet's superior performance in detoxifying memes compared to the other methods. These results underscore MemeDetoxNet's potential as an effective tool for content moderation on online platforms.

Self-Training Elicits Concise Reasoning in Large Language Models

Tergel Munkhbat, Namgyu Ho, Seo Hyun Kim, Yongjin Yang, Yujin Kim, Se-Young Yun

Chain-of-thought (CoT) reasoning has dramatically improved large language model performance on complex tasks. While recent "thinking" models, like OpenAI's o1, use substantial inference budgets to further boost performance, this raises concerns about inference efficiency due to potential redundancy. We hypothesize that LLMs have a latent capacity for more concise reasoning, evidenced by shorter, correct reasoning paths in their outputs. To elicit this capacity, we introduce fine-tuning methods leveraging self-generated data from Best-of-N sampling and few-shot conditioning. Our few-shot conditioned best-of-N sampling (FS-BoN) method significantly reduces reasoning length by 30%, or 2.4x more than previous baselines while maintaining accuracy. This implies fine-tuning with curated self-generated data can unlock latent concise reasoning, enabling more efficient inference.

From Conversation to Automation: Leveraging LLMs for Problem-Solving Therapy Analysis

Elham Aghakhani, Lu Wang, Karla T. Washington, George Demiris, Jina Huh-Yoo, Rezvaneh Rezapour

Problem-solving therapy (PST) is a structured psychological approach that helps individuals manage stress and resolve personal issues by guiding them through problem identification, solution brainstorming, decision-making, and outcome evaluation. As mental health care increasingly adopts technologies like chatbots and large language models (LLMs), it is important to thoroughly understand how each session of PST is conducted before attempting to auto mate it. We developed a comprehensive framework for PST annotation using established PST Core Strategies and a set of novel Facilitative Strategies to analyze a corpus of real-world therapy transcripts to determine which strategies are most prevalent. Using various LLM and transformer-based models, we found that GPT-4o outperformed all models, achieving the highest accuracy (0.76) in identifying all strategies. To gain deeper insights, we examined how strategies are applied by analyzing Therapeutic Dynamics (autonomy, self-disclosure, and metaphor), and linguistic patterns within our labeled data. Our research highlights LLMs' potential to automate therapy dialogue analysis, offering a scalable tool for mental health interventions. Our framework enhances PST by improving accessibility, effectiveness, and personalized support for therapists.

Almost AI, Almost Human: The Challenge of Detecting AI-Polished Writing

Shoumik Saha, Soheil Feizi

The growing use of large language models (LLMs) for text generation has led to widespread concerns about AI-generated content detection. However, an overlooked challenge is AI-polished text, where human-written content undergoes subtle refinements using AI tools. This raises a critical question: should minimally polished text be classified as AI-generated? Such classification can lead to false plagiarism accusations and misleading claims about AI prevalence in online content. In this study, we systematically evaluate eleven state-of-the-art AI-text detectors using our 'AI-Polished-Text Evaluation (APT-Eval)' dataset, which contains 15K samples refined at varying AI-involvement levels. Our findings reveal that detectors frequently flag even minimally polished text as AI-generated, struggle to differentiate between degrees of AI involvement, and exhibit biases against older and smaller models. These limitations highlight the urgent need for more nuanced detection methodologies.

Summary Factual Inconsistency Detection Based on LLMs Enhanced by Universal Information Extraction

Anguo Li, Lei Yu

Automatic text summarization has a potential flaw that affects the factuality of summaries. Recently, Large Language Models (LLMs) have been introduced as detectors for factual inconsistencies in summaries. However, LLM-based methods rely on reasoning capabilities and face challenges in terms of efficiency and explainability. We focus on decoupling LLMs' information extraction and reasoning capabilities to address prominent challenges, and propose a novel framework, UIEFID (Universal Information Extraction-enhanced Factual Inconsistency Detection). Our idea is to define a self-adaptive structured schema to guide fine-tuned LLMs in extracting unified structured information from documents and summaries, ultimately detecting the origins of inconsistencies in extraction information. The evaluation on 5 open-source models shows that UIEFID not only enhances the detection accuracy on the AGGREFACT benchmark but also significantly reduces redundant reasoning.

Red Queen: Exposing Latent Multi-Turn Risks in Large Language Models

Yifan Jiang, Kriti Aggarwal, Tanmay Laud, Kashif Munir, Jay Pujara, Subhabrata Mukherjee

The rapid advancement of large language models (LLMs) has unlocked diverse opportunities across domains and applications but has also raised concerns about their tendency to generate harmful responses under jailbreak attacks. However, most existing jailbreak strategies are single-turn with explicit malicious intent, failing to reflect the real-world scenario where interactions can be multi-turn and users can conceal their intents. Recent studies on Theory of Mind (ToM) reveal that LLMs often struggle to infer users' latent intent in such scenarios. Building on these limitations, we propose a novel jailbreak attack, RED QUEEN ATTACK, which constructs a multi-turn scenario, concealing the malicious intent under the guise of preventing harm. We generate 56k multi-turn concealment data points across 40 scenarios and 14 harmful categories, evaluating four LLM families of different sizes. Results show all models are vulnerable to RED QUEEN ATTACK, reaching 87.6% attack success rate (ASR) on GPT-4o and 77.1% on Llama3-70B. Compared to prior jailbreak attacks, the RED QUEEN ATTACK achieves superior performance on nine out of ten models, with ASR improvements ranging from 2% to 64%. Further analysis reveals that larger models exhibit greater vulnerability to our attack, primarily due to the combination of multi-turn structures and concealment strategies. To enhance safety, we propose RED QUEEN GUARD, a mitigation strategy reducing ASR to below 1% while maintaining model performance on standard benchmarks. We release all materials to support further research.

DiaLLM: EHR-Enhanced Clinical Conversational System for Clinical Test Recommendation and Diagnosis Prediction

Weijieying Ren, Tianxiang Zhao, Lei Wang, Tianchun Wang, Vasant G Honavar

Recent advances in Large Language Models (LLMs) have led to remarkable progresses in medical consultation. However, existing medical LLMs overlook the essential role of Electronic Health Records (EHR) and focus primarily on diagnosis recommendation, limiting their clinical applicability. We propose DiaLLM, the first medical LLM that integrates heterogeneous EHR data into clinically grounded dialogues, enabling clinical test recommendation, result interpretation, and diagnosis prediction to better align with real-world medical practice. To construct clinically grounded dialogues from EHR, we design a Clinical Test Reference (CTR) strategy that maps each clinical code to its corresponding description and classifies test results as "normal" or "abnormal". Additionally, DiaLLM employs a reinforcement learning framework for evidence acquisition and automated diagnosis. To handle the large action space, we introduce a reject sampling strategy to reduce redundancy and improve exploration efficiency. Furthermore, a confirmation reward and a class-sensitive diagnosis reward are designed to guide accurate diagnosis prediction. Extensive experimental results demonstrate that DiaLLM outperforms baselines in clinical test recommendation and diagnosis prediction. Our anonymous code is here\footnote{}.

Can Hallucination Correction Improve Video-Language Alignment?

Lingjun Zhao, Mingyang Xie, Paola Cascante-Bonilla, Hal Daumé III, Kwonjoon Lee

Large Vision-Language Models often generate hallucinated content that is not grounded in its visual inputs. While prior work focuses on mitigating hallucinations, we instead explore leveraging hallucination correction as a training objective to improve video-language alignment. We introduce HACA, a self-training framework learning to correct hallucinations in descriptions that do not align with the video content. By identifying and correcting inconsistencies, HACA enhances the model's ability to align video and textual representations for spatio-temporal reasoning. Our experimental results show consistent gains in video-caption binding and text-to-video retrieval tasks, demonstrating that hallucination correction-inspired tasks serve as an effective strategy for improving vision and language alignment.

Feature-Level Insights into Artificial Text Detection with Sparse Autoencoders

Kristian Kuznetsov, Laida Kushnareva, Anton Razzhigaev, Polina Druzhinina, Anastasia Voznyuk, Irina Piontkovskaya, Evgeny Burnaev, Serguei Barannikov

Artificial Text Detection (ATD) is becoming increasingly important with the rise of advanced Large Language Models (LLMs). Despite numerous efforts, no single algorithm performs consistently well across different types of unseen text or guarantees effective generalization to new LLMs. Interpretability plays a crucial role in achieving this goal. In this study, we enhance ATD interpretability by using Sparse Autoencoders (SAE) to extract features from Gemma-2-2b's residual stream. We identify both interpretable and efficient features, analyzing their semantics and relevance through domain- and model-specific statistics, a steering approach, and manual or LLM-based interpretation. Our methods offer valuable insights into how texts from various models differ from human-written content. We show that modern LLMs have a distinct writing style, especially in information-dense domains, even though they can produce human-like outputs with personalized prompts.

Benchmarking and Improving New Knowledge Acquisition during Continued Pre-training

Aochong Oliver Li, Tanya Goyal

Large language model (LLM) parameters store massive amounts of pre-training knowledge, but it is unclear how to continually update them to reflect changes in the real world. Both dataset and methodology works in existing literature focus on the entity-substitution framework which does not capture the breadth of complex real world updates. In this paper, we first introduce Knowledge Update Playground (KUP) to automatically construct a training corpus with realistic and complex updates to study this problems. Next, we introduce a novel approach, memory recall training (MRT), that trains the model to explicitly surface and condition on "recalled" knowledge both at train and test time. Our results show that the KUP benchmark is challenging for existing continued pre-training (CPT) approaches; LLMs after CPT still prefer outdated knowledge >40% of the time. Using our proposed MRT training approach significantly improves performance across datasets and evaluation settings (58.9 -> 71.0 for LLaMA-3.1-8B). Overall, we show that our KUP dataset, particularly an indirect probing evaluation setting, serves as a challenging testbed to facilitate future research.

CourtEval: A Courtroom-Based Multi-Agent Evaluation Framework

Sandeep Kumar, Abhijit A Nargund, Vivek Sridhar

Automated evaluation is crucial for assessing the quality of natural language text, especially in open-ended generation tasks, given the costly and time-consuming nature of human evaluation. Existing automatic evaluation metrics like ROUGE and BLEU often show low correlation with human judgments. As large language models (LLMs) continue to evolve, researchers have explored their use as alternatives to human evaluators. Although single-agent approaches have shown potential, results indicate that further progress is required to close the gap between their performance and the quality of human assessments. Acknowledging that human evaluations involve multiple annotators, the multi-agent approach allows LLMs to collaborate, enhancing efficiency and effectiveness in handling complex tasks. In this paper, we present CourtEval, a novel Multi-Agent Evaluation Framework modeled after courtroom dynamics. Each agent takes on a distinct role: the Grader, similar to a judge, assigns an initial score; the Critic, like a prosecutor, challenges this score; and the Defender, akin to a defense attorney, defends it. Based on the input from both the Critic and Defender, the Grader re-evaluates the score, leading to a more balanced and fair final decision through this adversarial process. CourtEval substantially outperforms the previous state-of-the-art methods in two meta-evaluation benchmarks in NLG evaluation, SummEval and TopicalChat.

Human Bias in the Face of AI: The Role of Human Judgement in AI Generated Text Evaluation

Tiffany Zhu, Iain Weissburg, Kexun Zhang, William Yang Wang

As AI advances in text generation, human trust in AI generated content remains constrained by biases that go beyond concerns of accuracy. This study explores how bias shapes the perception of AI versus human generated content. Through three experiments involving text rephrasing, news article summarization, and persuasive writing, we investigated how human raters respond to labeled and unlabeled content. While the raters could not differentiate the two types of texts in the blind test, they overwhelmingly favored content labeled as "Human Generated," over those labeled "AI Generated," by a preference score of over 30%. We observed the same pattern even when the labels were deliberately swapped. This human bias against AI has broader societal and cognitive implications, as it undervalues AI performance. This study highlights the limitations of human judgment in interacting with AI and offers a foundation for improving human-AI collaboration, especially in creative fields.

ADO: Automatic Data Optimization for Inputs in LLM Prompts

Sam Lin, Wenyue Hua, Lingyao Li, Zhenting Wang, Yongfeng Zhang

This study explores a novel approach to enhance the performance of Large Language Models (LLMs) through the optimization of input data within prompts. While previous research has primarily focused on refining instruction components and augmenting input data with in-context examples, our work investigates the potential benefits of optimizing the input data itself. We introduce a two-pronged strategy for input data optimization: content engineering and structural reformulation. Content engineering involves imputing missing values, removing irrelevant attributes, and enriching profiles by generating additional information inferred from existing attributes. Subsequent to content engineering, structural reformulation is applied to optimize the presentation of the modified content to LLMs, given their sensitivity to input format. Our findings suggest that these optimizations can significantly improve the performance of LLMs in various tasks, offering a promising avenue for future research in prompt engineering. The source code is available at https://github.com/glin2229/Automatic-Data-Optimization.

Do Vision-Language Models Have Internal World Models? Towards an Atomic Evaluation

Qiyue Gao, Xinyu Pi, Kevin Liu, Junrong Chen, Ruolan Yang, Xinqi Huang, Xinyu Fang, Lu Sun, Gautham Kishore, Bo Ai, Stone Tao, Mengyang Liu, Jiaxi Yang, Chao-Jung Lai, Chuanyang Jin, Jiannan Xiang, Benhao Huang, Zeming Chen, David Danks, Hao Su, Tianmin Shu, Ziqiao Ma, Lianhui Qin, Zhiting Hu

Internal world models (WMs) enable agents to understand the world's state and predict transitions, serving as the basis for advanced deliberative reasoning. Recent large Vision-Language Models (VLMs), such as GPT-4o and Gemini, exhibit potential as general-purpose WMs. While the latest studies have evaluated and shown limitations in specific capabilities such as visual understanding, a systematic evaluation of VLMs' fundamental WM abilities remains absent. Drawing on comparative psychology and cognitive science, we propose a two-stage framework that assesses Perception (visual, spatial, temporal, quantitative, and motion) and Prediction (mechanistic simulation, transitive inference, compositional inference) to provide an atomic evaluation of VLMs as WMs. Guided by this framework, we introduce WM-ABench, a large-scale benchmark comprising 23 fine-grained evaluation dimensions across 6 diverse simulated environments with controlled counterfactual simulations. Through 517 controlled experiments on 11 latest commercial and open-source VLMs, we find that these models exhibit striking limitations in basic world model abilities. For instance, all models perform at near-random accuracy when distinguishing motion trajectories. Additionally, they lack disentangled understanding---e.g., they tend to believe blue objects move faster than green ones. More rich results and analyses reveal significant gaps between VLMs and human-level world modeling.

IntentionESC: An Intention-Centered Framework for Enhancing Emotional Support in Dialogue Systems

Xinjie Zhang, Wenxuan Wang, Qin Jin

In emotional support conversations, unclear intentions can lead supporters to employ inappropriate strategies, inadvertently imposing their expectations or solutions on the seeker. Clearly defined intentions are essential for guiding both the supporter's motivations and the overall emotional support process. In this paper, we propose the Intention-centered Emotional Support Conversation (IntentionESC) framework, which defines the possible intentions of supporters in emotional support conversations, identifies key emotional state aspects for inferring these intentions, and maps them to appropriate support strategies. While Large Language Models (LLMs) excel in text generating, they fundamentally operate as probabilistic models trained on extensive datasets, lacking a true understanding of human thought processes and intentions. To address this limitation, we introduce the Intention CEntric Chain-of-Thought (ICECoT) mechanism. ICECoT enables LLMs to mimic human reasoning by analyzing emotional states, inferring intentions, and selecting suitable support strategies, thereby generating more effective emotional support responses. To train the model with ICECoT and integrate expert knowledge, we design an automated annotation pipeline that produces high-quality training data. Furthermore, we develop a comprehensive evaluation scheme to assess emotional support efficacy and conduct extensive experiments to validate our framework. Our data and code will be publically released to facilitate further research.

Rethinking Prompt-based Debiasing in Large Language Model

Xinyi Yang, Runzhe Zhan, Shu Yang, Junchao Wu, Lidia S. Chao, Derek F. Wong

Investigating bias in large language models (LLMs) is crucial for developing trustworthy AI. While prompt-based through prompt engineering is common, its effectiveness relies on the assumption that models inherently understand biases. Our study systematically analyzed this assumption using the BBQ and StereoSet benchmarks on both open-source models as well as commercial GPT model. Experimental results indicate that prompt-based is often superficial; for instance, the Llama2-7B-Chat model misclassified over 90% of unbiased content as biased, despite achieving high accuracy in identifying bias issues on the BBQ dataset. Additionally, specific evaluation and question settings in bias benchmarks often lead LLMs to choose "evasive answers", disregarding the core of the question and the relevance of the response to the context. Moreover, the apparent success of previous methods may stem from flawed evaluation metrics. Our research highlights a potential "false prosperity" in prompt-base efforts and emphasizes the need to rethink bias evaluation metrics to ensure truly trustworthy AI. We will release our data and code upon acceptance.

Training Language Model to Critique for Better Refinement

Tianshu Yu, Chao Xiang, mingchuan yang, Pei Ke, Bosi Wen, Cunxiang Wang, Jiale Cheng, Li Zhang, Xinyu Mu, Chuxiong Sun, Minlie Huang

Large language models (LLMs) have demonstrated remarkable evaluation and critique capabilities, providing insightful feedback and identifying flaws in various tasks. However, limited research has explored which types of critiques are most effective for improving model responses or how to generate such critiques. To address this gap, we introduce Refinement-oriented Critique Optimization (RCO), a novel framework designed to train critic models using refinement signals. RCO uses a feedback loop where critiques, generated by the critic model, guide the actor model in refining its responses. The critique utility (CU) quantifies the effectiveness of these refinements, serving as the reward signal for training the critic model. By focusing on critiques that lead to better refinements, RCO eliminates the need for direct critique preference assessment, ensuring that critiques driving meaningful improvements are rewarded. We evaluate RCO across five tasks—dialog generation, summarization, question answering, mathematical reasoning, and code generation—and show that it significantly outperforms traditional methods and open-source models in terms of critique quality and refinement outcomes. Our contributions include the introduction of RCO, a novel supervision scheme based on refined response preferences, and comprehensive experimental results that highlight the method's effectiveness in enhancing LLM critique-refinement loops. Code and data will be publicly available upon acceptance of this paper.

ProcrustesGPT: Compressing LLMs with Structured Matrices and Orthogonal Transformations

Ekaterina Grishina, Mikhail Gorbunov, Maxim Rakhuba

Large language models (LLMs) demonstrate impressive results in natural language processing tasks, but require a significant amount of computational and memory resources. Structured matrix representations is a promising way for reducing the number of parameters of these models. However, it seems unrealistic to expect that weight matrices of pretrained models can be accurately represented by structured matrices without any fine-tuning. To overcome this issue, we utilize the fact that LLM output is invariant under certain orthogonal transformations of weight matrices. This insight can be leveraged to identify transformations that significantly improve the compressibility of weights within the structured classes. The proposed approach is applicable to various types of structured matrices that support efficient projection operations.

10

Day 2: Tuesday 29 July

Oral Presentation

Session 7: Industry (Tuesday 29 July, 10:30-12:00, Hall L)

11:20 - 11:30

PlanGPT: Enhancing Urban Planning with a Tailored Agent Framework

He Zhu, Guanhua Chen, Wenjia Zhang

In the field of urban planning, general-purpose large language models often struggle to meet the specific needs of planners. Tasks like generating urban planning texts, retrieving related information, and evaluating planning documents pose unique challenges. To enhance the efficiency of urban professionals and overcome these obstacles, we introduce PlanGPT, the first specialized AI agent framework tailored for urban and spatial planning. Developed through collaborative efforts with professional urban planners, PlanGPT integrates a customized local database retrieval system, domain-specific knowledge activation capabilities, and advanced tool orchestration mechanisms. Through its comprehensive agent architecture, PlanGPT coordinates multiple specialized components to deliver intelligent assistance precisely tailored to the intricacies of urban planning workflows. Empirical tests demonstrate that PlanGPT framework has achieved advanced performance, providing comprehensive support that significantly enhances professional planning efficiency.

11:30 - 11:40

User Feedback Alignment for LLM-powered Exploration in Large-scale Recommendation Systems

Jianling Wang, Yifan Liu, Yinghao Sun, Xuejian Ma, Yueqi Wang, He Ma, Zhengyang Su, Minmin Chen, Mingyan Gao, Onkar Dalal, Ed H. Chi, Lichan Hong, Ningren Han, Haokai Lu

Exploration, the act of broadening user experiences beyond their established preferences, is challenging in large-scale recommendation systems due to feedback loops and limited signals on user exploration patterns. Large Language Models (LLMs) offer potential solutions by leveraging their world knowledge to recommend novel content outside these loops. A key challenge is aligning LLMs with user preferences while preserving their knowledge and reasoning. To enhance planning for new user interests using LLMs, this paper introduces a novel approach that combines hierarchical planning with LLM inference-time scaling. This method aims to improve recommendation relevancy without compromising novelty. We decouple novelty and user-alignment, training separate LLMs for each objective. We then scale up the novelty-focused LLM's inference and select the best-of-n predictions using the user-aligned LLM. Live experiments demonstrate efficacy, showing significant gains in both user satisfaction (measured by watch activity and active user counts) and exploration diversity.

11:30 - 11:40

Sentiment Reasoning for Healthcare

Khai-Nguyen Nguyen, Khai Le-Duc, Bach Phan Tat, Le Duy, Long Vo-Dang, Truong-Son Hy

Transparency in AI healthcare decision-making is crucial. By incorporating rationales to explain reason for each predicted label, users could understand Large Language Models (LLMs)'s reasoning to make better decision. In this work, we introduce a new task - Sentiment Reasoning - for both speech and text modalities, and our proposed multimodal multitask framework and the world's largest multimodal sentiment analysis dataset. Sentiment Reasoning is an auxiliary task in sentiment analysis where the model predicts both the sentiment label and generates the rationale behind it based on the input transcript. Our study conducted on both human transcripts and Automatic Speech Recognition (ASR) transcripts shows that Sentiment Reasoning helps improve model transparency by providing rationale for model prediction with quality semantically comparable to humans while also improving model's classification performance (+2% increase in both accuracy and macro- F1) via rationale-augmented fine-tuning. Also, no significant difference in the semantic quality of generated rationales between human and ASR transcripts. All code, data (five languages - Vietnamese, English, Chinese, German, and French) and models are published online.

11:40 - 11:50

SingaKids: A Multilingual Multimodal Dialogic Tutor for Language Learning

Zhengyuan Liu, Geyu Lin, Hui Li Tan, Huayun Zhang, Yanfeng Lu(luf@i2r.a-star.edu.sg), Xiaoxue Gao, Stella Xin Yin, He Sun, Hock Huan Goh, Lung Hsiang Wong, Nancy F. Chen

The integration of generative artificial intelligence into educational applications has enhanced personalized and interactive learning experiences, and it shows strong potential to promote young learners language acquisition. However, it is still challenging to ensure consistent and robust performance across different languages and cultural contexts, and kids-friendly design requires simplified instructions, engaging interactions, and age-appropriate scaffolding to maintain motivation and optimize learning outcomes. In this work, we introduce SingaKids, a dialogic tutor designed to facilitate language learning through picture description tasks. Our system integrates dense image captioning, multilingual dialogic interaction, speech understanding, and engaging speech generation to create an immersive learning environment in four languages: English, Mandarin, Malay, and Tamil. We further improve the system through multilingual pre-training, task-specific tuning, and scaffolding optimization. Empirical studies with elementary school students demonstrate that SingaKids provides effective dialogic teaching, benefiting learners at different performance levels.

Session 9: CL Journal (Tuesday 29 July, 14:00-15:30, Room 1.62)

15:17 - 15:28

LMLPA: Language Model Linguistic Personality Assessment

Lik-Hang Lee, Jingyao Zheng, Xian Wang, Simo Hosio, Xiaoxian Xu, Lik-Hang Lee

Large Language Models (LLMs) are increasingly used in everyday life and research. One of the most common use cases is conversational interactions, enabled by the language generation capabilities of LLMs. Just as between two humans, a conversation between an LLM-powered entity and a human depends on the personality of the conversants. However, measuring the personality of a given LLM is currently a challenge. This article introduces the Language Model Linguistic Personality Assessment (LMLPA), a system designed to evaluate the linguistic personalities of LLMs. Our system helps to understand LLMs' language generation capabilities by quantitatively assessing the distinct personality traits reflected in their linguistic outputs. Unlike traditional human-centric psychometrics, the LMLPA adapts a personality assessment questionnaire, specifically the Big Five Inventory, to align with the operational capabilities of LLMs, and also incorporates the findings from previous language-based personality measurement literature. To mitigate sensitivity to the order of options, our questionnaire is designed to be open-ended, resulting in textual answers. Thus, the Artificial Intelligence (AI) rater is needed to transform ambiguous personality information from text responses into clear numerical indicators of personality traits. Utilising Principal Component Analysis and reliability validation methods, our findings demonstrate that LLMs possess distinct personality traits that can be effectively quantified by the LMLPA. This research contributes to Human-Centered AI and Computational Linguistics, providing a robust framework for future studies to refine AI personality assessments and expand their applications in multiple areas, including education and manufacturing

Session 9: CL Journal (Tuesday 29 July, 14:00-15:30, Room 1.85)

15:06 - 15:17

Survey of Cultural Awareness in Language Models: Text and Beyond

Siddhesh Pawar Junyeong Park Jiho Jin Arnav Arora Junho Myung Srishti Yadav Faiz Ghifari Haznitrama Inhwa Song Alice Oh Isabelle Augenstein

Large-scale deployment of large language models (LLMs) in various applications, such as chatbots and virtual assistants, requires LLMs to be culturally sensitive to the user to ensure inclusivity. Culture has been widely studied in psychology and anthropology, and there has been a recent surge in research on making LLMs more culturally inclusive, going beyond multilinguality and building on findings from psychology and anthropology. In this paper, we survey efforts towards incorporating cultural awareness into text-based and multimodal LLMs. We start by defining cultural awareness in LLMs, taking definitions of culture from the anthropology and psychology literature as a point of departure. We then examine methodologies adopted for creating cross-cultural datasets, strategies for cultural inclusion in downstream tasks, and methodologies that have been used for benchmarking cultural awareness in LLMs. Further, we discuss the ethical implications of cultural alignment, the role of Human-Computer Interaction (HCI) in driving cultural inclusion in LLMs, and the role of cultural alignment in driving social science research. We finally provide pointers to future research based on our findings about gaps in the literature.

Session 9: Computational Social Science and Cultural Analytics (Tuesday 29 July, 14:00-15:30, Room 1.85)

14:00 - 14:11

SurveyPilot: an Agentic Framework for Automated Human Opinion Collection from Social Media

Viet Thanh Pham, Lizhen Qu, Zhuang Li, Suraj Sharma, Gholamreza Haffari

Viet Thanh Pham, Lizhen Qu, Zhuang Li, Suraj Sharma, Gholamreza Haffari

14:11 - 14:22

Investigating and Extending Homans' Social Exchange Theory with Large Language Model based Agents

Lei Wang, Zheqing Zhang, Xu Chen

Lei Wang, Zheqing Zhang, Xu Chen

14:22 - 14:33

Embracing Imperfection: Simulating Students with Diverse Cognitive Levels Using LLM-based Agents

Tao Wu, Jingyuan Chen, Wang Lin, Mengze Li, Yumeng Zhu, Ang Li, Kun Kuang, Fei Wu

Tao Wu, Jingyuan Chen, Wang Lin, Mengze Li, Yumeng Zhu, Ang Li, Kun Kuang, Fei Wu

14:33 - 14:44

CulFiT: A Fine-grained Cultural-aware LLM Training Paradigm via Multilingual Critique Data Synthesis

Ruixiang Feng, Shen Gao, Xiuying Chen, Lisi Chen, Shuo Shang

Ruixiang Feng, Shen Gao, Xiuying Chen, Lisi Chen, Shuo Shang

14:44 - 14:55

Multiple LLM Agents Debate for Equitable Cultural Alignment

Dayeon Ki, Rachel Rudinger, Tianyi Zhou, Marine Carpuat

Dayeon Ki, Rachel Rudinger, Tianyi Zhou, Marine Carpuat

14:55 - 15:06

NewsInterview: a Dataset and a Playground to Evaluate LLMs' Grounding Gap via Informational Interviews

Alexander Spangher, Michael Lu, Sriya Kalyan, Hyundong Justin Cho, Tenghao Huang, Weiyan Shi, Jonathan May

Alexander Spangher, Michael Lu, Sriya Kalyan, Hyundong Justin Cho, Tenghao Huang, Weiyan Shi, Jonathan May

Session 9: Dialogue and Interactive Systems (Tuesday 29 July, 14:00-15:30, Hall N.1)

14:00 - 14:11

CoMet: Metaphor-Driven Covert Communication for Multi-Agent Language Games

Shuhang Xu, Fangwei Zhong

Shuhang Xu, Fangwei Zhong

14:11 - 14:22

INTERACT: Enabling Interactive, Question-Driven Learning in Large Language Models

Aum Kendapadi, Kerem Zaman, Rakesh R Menon, Shashank Srivastava

Aum Kendapadi, Kerem Zaman, Rakesh R Menon, Shashank Srivastava

14:22 - 14:33

User-side Model Consistency Monitoring for Open Source Large Language Models Inference Services

Qijun Miao, Zhixuan Fang

Qijun Miao, Zhixuan Fang

14:33 - 14:44

PsyAdvisor: A Plug-and-Play Strategy Advice Planner with Proactive Questioning in Psychological Conversations

Yuxin Hu, Danni Liu, Bo Liu, Yida Chen, Jiuxin Cao, Yan Liu

Yuxin Hu, Danni Liu, Bo Liu, Yida Chen, Jiuxin Cao, Yan Liu

14:44 - 14:55

BOOKWORLD: From Novels to Interactive Agent Societies for Story Creation

Yiting Ran, Xintao Wang, Tian Qiu, Jiaqing Liang, Yanghua Xiao, Deqing Yang

Yiting Ran, Xintao Wang, Tian Qiu, Jiaqing Liang, Yanghua Xiao, Deqing Yang

15:06 - 15:17

Navigating Rifts in Human-LLM Grounding: Study and Benchmark

Omar Shaikh, Hussein Mozannar, Gagan Bansal, Adam Fourney, Eric Horvitz

Omar Shaikh, Hussein Mozannar, Gagan Bansal, Adam Fourney, Eric Horvitz

15:17 - 15:28

JailbreakRadar: Comprehensive Assessment of Jailbreak Attacks Against LLMs

Junjie Chu, Yugeng Liu, Ziqing Yang, Xinyue Shen, Michael Backes, Yang Zhang

Junjie Chu, Yugeng Liu, Ziqing Yang, Xinyue Shen, Michael Backes, Yang Zhang

Session 9: Discourse and Pragmatics (Tuesday 29 July, 14:00-15:30, Hall N.1)

14:55 - 15:06

Understanding Common Ground Misalignment in Goal-Oriented Dialog: A Case-Study with Ubuntu Chat Logs

Rupak Sarkar, Neha Srikanth, Taylor Hudson, Rachel Rudinger, Claire Bonial, Philip Resnik

Rupak Sarkar, Neha Srikanth, Taylor Hudson, Rachel Rudinger, Claire Bonial, Philip Resnik

Session 9: Efficient/Low-Resource Methods for NLP (Tuesday 29 July, 14:00-15:30, Hall M.1)

14:00 - 14:11

Towards Robust and Efficient Federated Low-Rank Adaptation with Heterogeneous Clients

Jabin Koo, Minwoo Jang, Jungseul Ok

Jabin Koo, Minwoo Jang, Jungseul Ok

14:11 - 14:22

Semantic Exploration with Adaptive Gating for Efficient Problem Solving with Language Models

Sungjae Lee, Hyejin Park, Jaechang Kim, Jungseul Ok

Sungjae Lee, Hyejin Park, Jaechang Kim, Jungseul Ok

14:22 - 14:33

Less for More: Enhanced Feedback-aligned Mixed LLMs for Molecule Caption Generation and Fine-Grained NLI Evaluation

Dimitris Gkoumas, Maria Liakata

Dimitris Gkoumas, Maria Liakata

14:33 - 14:44

Dually Self-Improved Counterfactual Data Augmentation Using Large Language Model

Luhao Zhang, Xinyu Zhang, Linmei Hu, Dandan Song, Liqiang Nie

Luhao Zhang, Xinyu Zhang, Linmei Hu, Dandan Song, Liqiang Nie

14:44 - 14:55

SPECTRA: Faster Large Language Model Inference with Optimized Internal and External Speculation

Nguyen-Khang Le, Truong Dinh Do, Le-Minh Nguyen

Nguyen-Khang Le, Truong Dinh Do, Le-Minh Nguyen

14:55 - 15:06

Diversity-oriented Data Augmentation with Large Language Models

Zaitian Wang, Jinghan Zhang, Xinhao Zhang, Kunpeng Liu, pengfei wang, Yuanchun Zhou

Zaitian Wang, Jinghan Zhang, Xinhao Zhang, Kunpeng Liu, pengfei wang, Yuanchun Zhou

15:06 - 15:17

GRaMPa: Subword Regularisation by Skewing Uniform Segmentation Distributions with an Efficient Path-counting Markov Model

Thomas Bauwens, David Kaczér, Miryam de Lhoneux

Thomas Bauwens, David Kaczér, Miryam de Lhoneux

15:17 - 15:28

Data-Constrained Synthesis of Training Data for De-Identification

Thomas Vakili, Aron Henriksson, Hercules Dalianis

Thomas Vakili, Aron Henriksson, Hercules Dalianis

Session 9: Industry (Tuesday 29 July, 14:00-15:30, Hall L)

14:00 - 14:10

Boosting Decision-Making Ability of LLMs with Speculative Reward Model Cost-Effectively

Jiawei Gu, Shangsong Liang

Effective decision-making in Large Language Models (LLMs) is essential for handling intricate tasks. However, existing approaches prioritize performance but often overlook the balance between effectiveness and computational cost. To address this, we first introduce the 3E Criteria to systematically assess the cost-effectiveness of search strategies, revealing that existing methods often trade significant efficiency for marginal performance gains. To improve LLM decision-making while maintaining efficiency, we propose the Speculative Reward Model (SRM), a plug-and-play framework that seamlessly integrates with existing search strategies. Specifically, SRM employs an external reward assigner to predict optimal actions, reducing reliance on LLMs' internal self-evaluation. And a speculative verification mechanism is used to prune suboptimal choices and guide the search toward more promising steps. We evaluate SRM on several complex decision-making tasks including mathematical reasoning, planning and numerical reasoning in specialized domains. Experimental results show that SRM reduces costs to 1/10 of the original search framework on average while maintaining effectiveness.

14:10 - 14:20

Arctic-TILT. Business Document Understanding at Sub-Billion Scale

Łukasz Borchmann, Michał Pietruszka, Wojciech Jaśkowski, Dawid Jurkiewicz, Piotr Halama, Paweł Józiak, Łukasz Garncarek, Paweł Liskowski, Karolina Szyndler, Andrzej Gretkowski, Julita Ołtusek, Gabriela Nowakowska, Artur Zawłocki, Łukasz Duhr, Paweł Dyda, Michał Turski

The vast portion of workloads employing LLMs involves answering questions grounded on PDF or scanned content. We introduce the Arctic-TILT achieving accuracy on par with models 1000\times its size on these use cases. It can be finetuned and deployed on a single 24GB GPU, lowering operational costs while processing rich documents with up to 400k tokens. The model establishes state-of-the-art results on seven diverse Document Understanding benchmarks, as well as provides reliable confidence scores and quick inference, essential for processing files in large-scale or time-sensitive enterprise environments. We release Arctic-TILT weights and an efficient vLLM-based implementation on a permissive license.

14:20 - 14:30

TablePilot: Recommending Human-Preferred Tabular Data Analysis with Large Language Models

Deyin Yi, Yihao Liu, Lang Cao, Mengyu Zhou, Haoyu Dong, Shi Han, Dongmei Zhang

Tabular data analysis is crucial in many scenarios, yet efficiently identifying relevant queries and results for new tables remains challenging due to data complexity, diverse analytical operations, and high-quality analysis requirements. To address these challenges, we aim to recommend query–code–result triplets tailored for new tables in tabular data analysis workflows. In this paper, we present TablePilot, a pioneering tabular data analysis framework leveraging large language models to autonomously generate comprehensive and superior analytical results without relying on user profiles or prior interactions. Additionally, we propose Rec-Align, a novel method to further improve recommendation quality and better align with human preferences. Experiments on DART, a dataset specifically designed for comprehensive tabular data analysis recommendation, demonstrate the effectiveness of our framework. Based on GPT-4o, the tuned TablePilot achieves 77.0% top-5 recommendation recall. Human evaluations further highlight its effectiveness in optimizing tabular data analysis workflows.

14:30 - 14:40

Model Merging for Knowledge Editing

Zichuan Fu, Xian Wu, Guojing Li, Yingying Zhang, Yefeng Zheng, Tianshi Ming, Yejing Wang, Wanyu Wang, Xiangyu Zhao

Large Language Models (LLMs) require continuous updates to maintain accurate and current knowledge as the world evolves. While existing knowledge editing approaches offer various solutions for knowledge updating, they often struggle with sequential editing scenarios and harm the general capabilities of the model, thereby significantly hampering their practical applicability. This paper proposes a two-stage framework combining robust supervised fine-tuning (R-SFT) with model merging for knowledge editing. Our method first fine-tunes the LLM to internalize new knowledge fully, then merges the fine-tuned model with the original foundation model to preserve newly acquired knowledge and general capabilities. Experimental results demonstrate that our approach significantly outperforms existing methods in sequential editing while better preserving the original performance of the model, all without requiring any architectural changes. Code is available at [Applied-Machine-Learning-Lab/MM4KE](https://github.com/Applied-Machine-Learning-Lab/MM4KE).

14:40 - 14:50

Conceptual Diagnostics for Knowledge Graphs and Large Language Models

Rosario Uceda Sosa, Maria Chang, Karthikeyan Natesan Ramamurthy, Moninder Singh

Industrial applications pose heightened requirements for consistency and reliability of large language models (LLMs). While LLMs are being tested with increasingly complex reasoning tasks, we argue that much can be learned via diagnostic tools that probe a fundamentally basic type of reasoning: conceptual consistency, e.g., a rule applying to "all surgeons" must also apply to "cardiac surgeons" since a cardiac surgeon is a type of surgeon. In this emerging industry track submission, we propose a method that takes concept hierarchies from a knowledge graph (KG) and automatically generates benchmarks that test conceptual consistency in LLMs. We develop a multi-domain benchmark that reveals rates of conceptual inconsistencies in several state of the art LLMs. Additionally, we use measured levels of inconsistency and disagreement in LLMs to find potentially problematic subgraphs in the reference KG. As such, it offers a scalable complement to symbolic curation, maintenance, and refinement of knowledge graphs, which is a critical activity in KG-based industrial applications.

14:50 - 15:00

Optimization before Evaluation: Evaluation with Unoptimized Prompts Can be Misleading

Nicholas Sadjoli, Tim Siefken, Atin Ghosh, Yifan Mai, Daniel Dahlmeier

Current Large Language Model (LLM) evaluation frameworks utilize the same static prompt template across all models under evaluation. This differs from the common industry practice of using prompt optimization (PO) techniques to optimize the prompt for each model to maximize application performance. In this paper, we investigate the effect of PO towards LLM evaluations. Our results on public academic and internal industry benchmarks show that PO greatly affects the final ranking of models. This highlights the importance of practitioners performing PO per model when conducting evaluations to choose the best model for a given task.

15:00 - 15:10

Learning from Litigation: Graphs for Retrieval and Reasoning in eDiscovery

Sounak Lahiri, Sumit Pai, Tim Weninger, Sanmitra Bhattacharya

Electronic Discovery (eDiscovery) requires identifying relevant documents from vast collections for legal production requests. While artificial intelligence (AI) and natural language processing (NLP) have improved document review efficiency, current methods still struggle with legal entities, citations, and complex legal artifacts. To address these challenges, we introduce DISCOvery Graph (DISCOG), an emerging system that integrates knowledge graphs for enhanced document ranking and classification, augmented by LLM-driven reasoning. DISCOG outperforms strong baselines in F1-score, precision, and recall across both balanced and imbalanced datasets. In real-world deployments, it has reduced litigation-related document review costs by approximately 98%, demonstrating significant business impact.

15:10 - 15:20

Reinforcement Learning for Adversarial Query Generation to Enhance Relevance in Cold-Start Product Search

Akshay Jagatap, Neeraj Anand, Sonali Singh, Prakash Mandayam Comar

Accurate mapping of queries to product categories is crucial for efficient retrieval and ranking of relevant products in e-commerce search. Conventionally, such query classification models rely on supervised learning using historical user interactions, but their effectiveness diminishes in cold-start scenarios, where new categories or products lack sufficient training data. This results in poor query-to-category mappings, negatively affecting retrieval and ranking. Synthetic query generation has emerged as a promising solution by augmenting training data; however, existing methods do not incorporate feedback from the query relevance model, limiting their ability to generate queries that enhance product retrieval. To address this, we propose an adversarial reinforcement learning framework that optimizes an LLM-based generator to expose weaknesses in query classification models. The generator produces synthetic queries to augment the classifier's training set, ultimately improving its performance. Additionally, we introduce a structured reward signal to ensure stable training. Experiments on public datasets show an average PR-AUC improvement of +1.82\% on benchmarks and +3.26\% on a proprietary dataset, demonstrating the framework's effectiveness in enhancing query classification and mitigating cold-start challenges.

15:20 - 15:30

Auto Review: Second Stage Error Detection for Highly Accurate Information Extraction from Phone Conversations

Ayesha Qamar, Arushi Raghuvanshi, Conal Sathi, Youngseo Son

Automating benefit verification phone calls saves time in healthcare and helps patients receive treatment faster. It is critical to obtain highly accurate information in these phone calls, as it can affect a patient's healthcare journey. Given the noise in phone call transcripts, we have a two-stage system that involves a post-call review phase for potentially noisy fields, where human reviewers manually verify the extracted data—a labor-intensive task. To automate this stage, we introduce Auto Review, which significantly reduces manual effort while maintaining a high bar for accuracy. This system, being highly reliant on call transcripts, suffers a performance bottleneck due to automatic speech recognition (ASR) issues. This problem is further exacerbated by the use of domain-specific jargon in the calls. In this work, we propose a second-stage postprocessing pipeline for accurate information extraction. We improve accuracy by using multiple ASR alternatives and a pseudo-labeling approach that does not require manually corrected transcripts. Experiments with general-purpose large language models and feature-based model pipelines demonstrate substantial improvements in the quality of corrected call transcripts, thereby enhancing the efficiency of Auto Review.

Session 9: Information Extraction (Tuesday 29 July, 14:00-15:30, Hall N.2)

14:00 - 14:11

Cuckoo: An IE Free Rider Hatched by Massive Nutrition in LLM's Nest

Letian Peng, Zilong Wang, Feng Yao, Jingbo Shang

Letian Peng, Zilong Wang, Feng Yao, Jingbo Shang

14:22 - 14:33

Revisiting Classical Chinese Event Extraction with Ancient Literature Information

Xiaoyi Bao, Zhongqing Wang, Jinghang Gu, Chu-Ren Huang

Xiaoyi Bao, Zhongqing Wang, Jinghang Gu, Chu-Ren Huang

14:44 - 14:55

Evaluating Design Decisions for Dual Encoder-based Entity Disambiguation

Susanna Rücker, Alan Akbik

Susanna Rücker, Alan Akbik

15:06 - 15:17

Synergizing Unsupervised Episode Detection with LLMs for Large-Scale News Events

Priyanka Kargupta, Yunyi Zhang, Yizhu Jiao, Siru Ouyang, Jiawei Han

Priyanka Kargupta, Yunyi Zhang, Yizhu Jiao, Siru Ouyang, Jiawei Han

15:17 - 15:28

Map\&Make: Schema Guided Text to Table Generation

Naman Ahuja, Fenil Bardoliya, Chitta Baral, Vivek Gupta

Naman Ahuja, Fenil Bardoliya, Chitta Baral, Vivek Gupta

Session 9: Interpretability and Analysis of Models for NLP (Tuesday 29 July, 14:00-15:30, Hall M.2)

14:00 - 14:11

Unraveling the Mechanics of Learning-Based Demonstration Selection for In-Context Learning
Hui Liu, Wenya Wang, Hao Sun, Chris XING TIAN, Chenqi Kong, Xin Dong, Haoliang Li
Hui Liu, Wenya Wang, Hao Sun, Chris XING TIAN, Chenqi Kong, Xin Dong, Haoliang Li

14:11 - 14:22

Position-aware Automatic Circuit Discovery
Tal Haklay, Hadas Orgad, David Bau, Aaron Mueller, Yonatan Belinkov
Tal Haklay, Hadas Orgad, David Bau, Aaron Mueller, Yonatan Belinkov

14:22 - 14:33

Your Model is Overconfident, and Other Lies We Tell Ourselves
Timothee Mickus, Aman Sinha, Raúl Vázquez
Timothee Mickus, Aman Sinha, Raúl Vázquez

14:33 - 14:44

LLM-Guided Semantic-Aware Clustering for Topic Modeling
Jianghan Liu, Ziyu Shang, Wenjun Ke, Peng Wang, Zhizhao Luo, Jiajun Liu, Guozheng Li, Yining Li
Jianghan Liu, Ziyu Shang, Wenjun Ke, Peng Wang, Zhizhao Luo, Jiajun Liu, Guozheng Li, Yining Li

14:44 - 14:55

Towards Harmonized Uncertainty Estimation for Large Language Models
Rui Li, Jing Long, Muge Qi, Heming Xia, Lei Sha, Peiyi Wang, Zhifang Sui
Rui Li, Jing Long, Muge Qi, Heming Xia, Lei Sha, Peiyi Wang, Zhifang Sui

14:55 - 15:06

Beyond Prompt Engineering: Robust Behavior Control in LLMs via Steering Target Atoms
Mengru Wang, Ziwen Xu, Shengyu Mao, Shumin Deng, Zhaopeng Tu, Huajun Chen, Ningyu Zhang
Mengru Wang, Ziwen Xu, Shengyu Mao, Shumin Deng, Zhaopeng Tu, Huajun Chen, Ningyu Zhang

15:06 - 15:17

What Happened in LLMs Layers when Trained for Fast vs. Slow Thinking: A Gradient Perspective
Ming Li, Yanhong Li, Tianyi Zhou
Ming Li, Yanhong Li, Tianyi Zhou

15:17 - 15:28

Multi-Level Explanations for Generative Language Models
Lucas Monteiro Paes, Dennis Wei, Hyo Jin Do, Hendrik Strobelt, Ronny Luss, Amit Dhurandhar, Manish Nagireddy, Karthikeyan Natesan Ramamurthy, Prasanna Sattigeri, Werner Geyer, Soumya Ghosh
Lucas Monteiro Paes, Dennis Wei, Hyo Jin Do, Hendrik Strobelt, Ronny Luss, Amit Dhurandhar, Manish Nagireddy, Karthikeyan Natesan Ramamurthy, Prasanna Sattigeri, Werner Geyer, Soumya Ghosh

Session 9: Language Modeling (Tuesday 29 July, 14:00-15:30, Hall C)

14:00 - 14:11

Jailbreak Large Vision-Language Models Through Multi-Modal Linkage

Yu Wang, Xiaofei Zhou, Yichen Wang, Geyuan Zhang, Tianxing He

Yu Wang, Xiaofei Zhou, Yichen Wang, Geyuan Zhang, Tianxing He

14:11 - 14:22

TokAlign: Efficient Vocabulary Adaptation via Token Alignment

Chong Li, Jiajun Zhang, Chengqing Zong

Chong Li, Jiajun Zhang, Chengqing Zong

14:22 - 14:33

Byte Latent Transformer: Patches Scale Better Than Tokens

Artidoro Pagnoni, Ramakanth Pasunuru, Pedro Rodriguez, John Nguyen, Benjamin Muller, Margaret Li, Chunting Zhou, LILI YU, Jason E Weston, Luke Zettlemoyer, Gargi Ghosh, Mike Lewis, Ari Holtzman, Srini Iyer

Artidoro Pagnoni, Ramakanth Pasunuru, Pedro Rodriguez, John Nguyen, Benjamin Muller, Margaret Li, Chunting Zhou, LILI YU, Jason E Weston, Luke Zettlemoyer, Gargi Ghosh, Mike Lewis, Ari Holtzman, Srini Iyer

14:33 - 14:44

Cooperative or Competitive? Understanding the Interaction between Attention Heads From A Game Theory Perspective

Xiaoye Qu, Zengqi Yu, Dongrui Liu, Wei Wei, Daizong Liu, Jianfeng Dong, Yu Cheng

Xiaoye Qu, Zengqi Yu, Dongrui Liu, Wei Wei, Daizong Liu, Jianfeng Dong, Yu Cheng

14:44 - 14:55

Native Sparse Attention: Hardware-Aligned and Natively Trainable Sparse Attention

Jingyang Yuan, Huazuo Gao, Damai Dai, Junyu Luo, Liang Zhao, Zhengyan Zhang, Zhenda Xie, Yuxing Wei, Lean Wang, Zhiping Xiao, Yuqing Wang, Chong Ruan, Ming Zhang, Wenfeng Liang, Wangding Zeng

Jingyang Yuan, Huazuo Gao, Damai Dai, Junyu Luo, Liang Zhao, Zhengyan Zhang, Zhenda Xie, Yuxing Wei, Lean Wang, Zhiping Xiao, Yuqing Wang, Chong Ruan, Ming Zhang, Wenfeng Liang, Wangding Zeng

14:55 - 15:06

Understanding Language Model Scaling Laws in Terms of Training Dynamics via Loss Deceleration and Zero-Sum Learning

Andrei Mircea, Ekaterina Lobacheva, Supriyo Chakraborty, Nima Chitsazan, Irina Rish

Andrei Mircea, Ekaterina Lobacheva, Supriyo Chakraborty, Nima Chitsazan, Irina Rish

15:06 - 15:17

Unveiling the Potential of BERT-family: A New Recipe for Building Scalable, General and Competitive Large Language Models

Yisheng Xiao, Juntao Li, Wenpeng Hu, Min Zhang

Yisheng Xiao, Juntao Li, Wenpeng Hu, Min Zhang

15:17 - 15:28

OpenCoder: The Open Cookbook for Top-Tier Code Large Language Models

Siming Huang, Tianhao Cheng, Jason Klein Liu, Weidi Xu, JIARAN HAO, Liuyihan Song, Yang Xu, Jian Yang, Jiaheng Liu, Chenchen Zhang, Linzheng Chai, Ruifeng Yuan, Xianzhen Luo, Qiufeng Wang, YuanTao Fan, Qingfu Zhu, Zhaoxiang Zhang, Yang Gao, Jie Fu, Qian Liu, Houyi Li, Ge Zhang, Yuan Qi, Xu Yinghui, Wei Chu, Zili Wang

Siming Huang, Tianhao Cheng, Jason Klein Liu, Weidi Xu, JIARAN HAO, Liuyihan Song, Yang Xu, Jian Yang, Jiaheng Liu, Chenchen Zhang, Linzheng Chai, Ruifeng Yuan, Xianzhen Luo, Qiufeng Wang, YuanTao Fan, Qingfu Zhu, Zhaoxiang Zhang, Yang Gao, Jie Fu, Qian Liu, Houyi Li, Ge Zhang, Yuan Qi, Xu Yinghui, Wei Chu, Zili Wang

Session 9: Multilinguality and Language Diversity (Tuesday 29 July, 14:00-15:30, Room 1.15-16)

14:00 - 14:11

Do Large Language Models have an English Accent? Evaluating and Improving the Naturalness of Multilingual LLMs

Yanzhu Guo, Simone Conia, Zelin Zhou, Min Li, Saloni Potdar, Henry Xiao

Yanzhu Guo, Simone Conia, Zelin Zhou, Min Li, Saloni Potdar, Henry Xiao

14:11 - 14:22

Data Quality Issues in Multilingual Speech Datasets: The Need for Sociolinguistic Awareness and Proactive Language Planning

Mingfei Lau, Qian Chen, Yeming Fang, Tingting Xu, Tongzhou Chen, Pavel Golik

Mingfei Lau, Qian Chen, Yeming Fang, Tingting Xu, Tongzhou Chen, Pavel Golik

14:22 - 14:33

INJONGO: A Multicultural Intent Detection and Slot-filling Dataset for 16 African Languages

Hao Yu, Jesujoba Oluwadara Alabi, Andiswa Bukula, Jian Yun Zhuang, En-Shiun Annie Lee, Tadesse Kebede Guge, Israel Abebe Azime, Happy Buzaaba, Blessing Kudzaishe Sibanda, Godson Koffi KALIPE, Jonathan Mukiibi, Salomon KABONGO KABENAMUALU, Mmasibidi Setaka, Lolwethu Ndolela, Nkiruka Odu, Rooweither Mabuya, Shamsuddeen Hassan Muhammad, Salomey Osei, Sokhar Samb, Dietrich Klakow, David Ifeoluwa Adelani

Hao Yu, Jesujoba Oluwadara Alabi, Andiswa Bukula, Jian Yun Zhuang, En-Shiun Annie Lee, Tadesse Kebede Guge, Israel Abebe Azime, Happy Buzaaba, Blessing Kudzaishe Sibanda, Godson Koffi KALIPE, Jonathan Mukiibi, Salomon KABONGO KABENAMUALU, Mmasibidi Setaka, Lolwethu Ndolela, Nkiruka Odu, Rooweither Mabuya, Shamsuddeen Hassan Muhammad, Salomey Osei, Sokhar Samb, Dietrich Klakow, David Ifeoluwa Adelani

14:33 - 14:44

Cross-Lingual Optimization for Language Transfer in Large Language Models

Jungseob Lee, Seongtae Hong, Hyeonseok Moon, Heuiseok Lim

Jungseob Lee, Seongtae Hong, Hyeonseok Moon, Heuiseok Lim

14:44 - 14:55

BMIKE-53: Investigating Cross-Lingual Knowledge Editing with In-Context Learning

Ercong Nie, Bo Shao, Mingyang Wang, Zifeng Ding, Helmut Schmid, Hinrich Schuetze

Ercong Nie, Bo Shao, Mingyang Wang, Zifeng Ding, Helmut Schmid, Hinrich Schuetze

14:55 - 15:06

From Isolates to Families: Using Neural Networks for Automated Language Affiliation

Frederic Blum, Steffen Herbold, Johann-Mattis List

Frederic Blum, Steffen Herbold, Johann-Mattis List

15:06 - 15:17

Identifying Open Challenges in Language Identification

Rob van der Goot

Rob van der Goot

15:17 - 15:28

Training-free LLM Merging for Multi-task Learning

Zichuan Fu, Xian Wu, Yejing Wang, Wanyu Wang, Shanshan Ye, Hongzhi Yin, Yi Chang, Yefeng Zheng, Xiangyu Zhao

Zichuan Fu, Xian Wu, Yejing Wang, Wanyu Wang, Shanshan Ye, Hongzhi Yin, Yi Chang, Yefeng Zheng, Xiangyu Zhao

Session 9: Phonology, Morphology, and Word Segmentation AND Syntax: Tagging, Chunking and Parsing (Tuesday 29 July, 14:00-15:30, Room 1.86)

14:00 - 14:11

A Systematic Study of Compositional Syntactic Transformer Language Models

Yida Zhao, Hao Xve, Xiang Hu, Kewei Tu

Yida Zhao, Hao Xve, Xiang Hu, Kewei Tu

14:11 - 14:22

An Effective Incorporating Heterogeneous Knowledge Curriculum Learning for Sequence Labeling

Xuemei Tang, Jun Wang, Qi Su, Chu-Ren Huang, Jinghang Gu

Xuemei Tang, Jun Wang, Qi Su, Chu-Ren Huang, Jinghang Gu

14:22 - 14:33

Dynamic Head Selection for Neural Lexicalized Constituency Parsing

Yang Hou, Zhenghua Li

Yang Hou, Zhenghua Li

14:33 - 14:44

Segment First or Comprehend First? Explore the Limit of Unsupervised Word Segmentation with Large Language Models

Zihong Zhang, Liqi He, Zuchao Li, Lefei Zhang, hai zhao, Bo Du

Zihong Zhang, Liqi He, Zuchao Li, Lefei Zhang, hai zhao, Bo Du

14:44 - 14:55

Hierarchical Bracketing Encodings for Dependency Parsing as Tagging

Ana Ezquerro, David Vilares, Anssi Yli-Jyrä, Carlos Gómez-Rodríguez

Ana Ezquerro, David Vilares, Anssi Yli-Jyrä, Carlos Gómez-Rodríguez

14:55 - 15:06

Chinese Inertial GAN for Handwriting Signal Generation and Recognition

Yifeng Wang, Yi Zhao

Yifeng Wang, Yi Zhao

15:06 - 15:17

Bregman Conditional Random Fields: Sequence Labeling with Parallelizable Inference Algorithms

Caio Corro, Mathieu Lacroix, Joseph Le Roux

Caio Corro, Mathieu Lacroix, Joseph Le Roux

15:17 - 15:28

Programming by Example meets Historical Linguistics: A Large Language Model Based Approach to Sound Law Induction

Atharva Naik, Darsh Agrawal, Hong Sng, Clayton Marr, Kexun Zhang, Nathaniel Romney Robinson, Kalvin Chang, Rebecca Byrnes, Aravind Mysore, Carolyn Rose, David R Mortensen

Atharva Naik, Darsh Agrawal, Hong Sng, Clayton Marr, Kexun Zhang, Nathaniel Romney Robinson, Kalvin Chang, Rebecca Byrnes, Aravind Mysore, Carolyn Rose, David R Mortensen

Session 9: Resources and Evaluation (Tuesday 29 July, 14:00-15:30, Hall B)

14:00 - 14:11

BRIGHTER: BRIdging the Gap in Human-Annotated Textual Emotion Recognition Datasets for 28 Languages

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, Guanqun Cao, Rodrigo Tufiño, Rendi Chevi, Chiamaka Ijeoma Chukwuneke, Alexandra Ciobotaru, Daryna Dementieva, Murja Sani Gadanya, Robert Geislinger, Bela Gipp, Oumaima Hourrane, Oana Ignat, Falalu Ibrahim Lawan, Rooweither Mabuya, Rahmad Mahendra, Vukosi Marivate, Alexander Panchenko, Andrew Piper, Charles Henrique Porto Ferreira, Vitaly Protasov, Samuel Rutunda, Manish Shrivastava, Aura Cristina Udrea, Lilian Diana Awuor Wanzare, Sophie Wu, Florian Valentin Wunderlich, Hanif Muhammad Zhafran, Tianhui Zhang, Yi Zhou, Saif M. Mohammad

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, Guanqun Cao, Rodrigo Tufiño, Rendi Chevi, Chiamaka Ijeoma Chukwuneke, Alexandra Ciobotaru, Daryna Dementieva, Murja Sani Gadanya, Robert Geislinger, Bela Gipp, Oumaima Hourrane, Oana Ignat, Falalu Ibrahim Lawan, Rooweither Mabuya, Rahmad Mahendra, Vukosi Marivate, Alexander Panchenko, Andrew Piper, Charles Henrique Porto Ferreira, Vitaly Protasov, Samuel Rutunda, Manish Shrivastava, Aura Cristina Udrea, Lilian Diana Awuor Wanzare, Sophie Wu, Florian Valentin Wunderlich, Hanif Muhammad Zhafran, Tianhui Zhang, Yi Zhou, Saif M. Mohammad

14:11 - 14:22

The Alternative Annotator Test for LLM-as-a-Judge: How to Statistically Justify Replacing Human Annotators with LLMs

Nitay Calderon, Roi Reichart, Rotem Dror

Nitay Calderon, Roi Reichart, Rotem Dror

14:22 - 14:33

AntiLeakBench: Preventing Data Contamination by Automatically Constructing Benchmarks with Updated Real-World Knowledge

Xiaobao Wu, Liangming Pan, Yuxi Xie, Ruiwen Zhou, Shuai Zhao, Yubo Ma, Mingzhe Du, Rui Mao, Anh Tuan Luu, William Yang Wang

Xiaobao Wu, Liangming Pan, Yuxi Xie, Ruiwen Zhou, Shuai Zhao, Yubo Ma, Mingzhe Du, Rui Mao, Anh Tuan Luu, William Yang Wang

14:33 - 14:44

FactBench: A Dynamic Benchmark for In-the-Wild Language Model Factuality Evaluation

Farima Fatahi Bayat, Lechen Zhang, Sheza Munir, Lu Wang

Farima Fatahi Bayat, Lechen Zhang, Sheza Munir, Lu Wang

14:44 - 14:55

Data Laundering: Artificially Boosting Benchmark Results through Knowledge Distillation

Jonibek Mansurov, Akhmed Sakip, Alham Fikri Aji

Jonibek Mansurov, Akhmed Sakip, Alham Fikri Aji

14:55 - 15:06

FineReason: Evaluating and Improving LLMs' Deliberate Reasoning through Reflective Puzzle Solving

Session 9: Sentiment Analysis, Stylistic Analysis, and Argument Mining (Tuesday 29 July, 14:00-15:30, Room 1.62)

14:00 - 14:11

Mining Complex Patterns of Argumentative Reasoning in Natural Language Dialogue

Ramon Ruiz-Dolz, Zlata Kikteva, John Lawrence

Ramon Ruiz-Dolz, Zlata Kikteva, John Lawrence

14:11 - 14:22

A Multi-persona Framework for Argument Quality Assessment

Bojun Jin, Jianzhu Bao, Yufang Hou, Yang Sun, Yice Zhang, Huajie Wang, Bin Liang, Ruifeng Xu

Bojun Jin, Jianzhu Bao, Yufang Hou, Yang Sun, Yice Zhang, Huajie Wang, Bin Liang, Ruifeng Xu

14:22 - 14:33

Towards Comprehensive Argument Analysis in Education: Dataset, Tasks, and Method

Yupei Ren, Xinyi Zhou, Ning Zhang, Shangqing Zhao, Man Lan, Xiaopeng Bai

Yupei Ren, Xinyi Zhou, Ning Zhang, Shangqing Zhao, Man Lan, Xiaopeng Bai

14:33 - 14:44

DS^2-ABSA: Dual-Stream Data Synthesis with Label Refinement for Few-Shot Aspect-Based Sentiment Analysis

Hongling Xu, Yice Zhang, Qianlong Wang, Ruifeng Xu

Hongling Xu, Yice Zhang, Qianlong Wang, Ruifeng Xu

14:44 - 14:55

Sheep's Skin, Wolf's Deeds: Are LLMs Ready for Metaphorical Implicit Hate Speech?

Jingjie Zeng, Yuanyuan Sun, zekun wang, Liang Yang, Hongfei Lin

Jingjie Zeng, Yuanyuan Sun, zekun wang, Liang Yang, Hongfei Lin

14:55 - 15:06

It's Not Bragging If You Can Back It Up: Can LLMs Understand Braggings?

Jingjie Zeng, Huayang Li, Yuanyuan Sun, Liang Yang, Hongfei Lin

Jingjie Zeng, Huayang Li, Yuanyuan Sun, Liang Yang, Hongfei Lin

15:06 - 15:17

Classifying Unreliable Narrators with Large Language Models

Anneliese Brei, Katharine Henry, Abhisheik Sharma, Shashank Srivastava, Snigdha Chaturvedi

Anneliese Brei, Katharine Henry, Abhisheik Sharma, Shashank Srivastava, Snigdha Chaturvedi

Session 9: Speech recognition, text-to-speech and spoken language understanding (Tuesday 29 July, 14:00-15:30, Room 1.61)

14:00 - 14:11

Takin-VC: Expressive Zero-Shot Voice Conversion via Adaptive Hybrid Content Encoding and Enhanced Timbre Modeling

Yang Yuguang, Yu Pan, Jixun Yao, xiang zhang, Jianhao Ye, Hongbin Zhou, Lei Xie, Lei Ma, Jianjun Zhao

Yang Yuguang, Yu Pan, Jixun Yao, xiang zhang, Jianhao Ye, Hongbin Zhou, Lei Xie, Lei Ma, Jianjun Zhao

14:11 - 14:22

Distilling an End-to-End Voice Assistant Without Instruction Training Data

William Barr Held, Yanzhe Zhang, Weiyan Shi, Minzhi Li, Michael J Ryan, Diyi Yang

William Barr Held, Yanzhe Zhang, Weiyan Shi, Minzhi Li, Michael J Ryan, Diyi Yang

14:22 - 14:33

Mind the Gap: Static and Interactive Evaluations of Large Audio Models

Minzhi Li, William Barr Held, Michael J Ryan, Kunat Pipatanakul, Potsawee Manakul, Hao Zhu, Diyi Yang

Minzhi Li, William Barr Held, Michael J Ryan, Kunat Pipatanakul, Potsawee Manakul, Hao Zhu, Diyi Yang

14:33 - 14:44

WavRAG: Audio-Integrated Retrieval Augmented Generation for Spoken Dialogue Models

Yifu Chen, Shengpeng Ji, Haoxiao Wang, Ziqing Wang, Siyu Chen, Jinzheng He, Jin Xu, Zhou Zhao

Yifu Chen, Shengpeng Ji, Haoxiao Wang, Ziqing Wang, Siyu Chen, Jinzheng He, Jin Xu, Zhou Zhao

14:44 - 14:55

Bridging Discrete Codec Representations and Speech Language Models
Shengpeng Ji, Minghui Fang, Jialong Zuo, Ziyue Jiang, Dingdong WANG, Hanting Wang, Hai Huang, Zhou Zhao
Shengpeng Ji, Minghui Fang, Jialong Zuo, Ziyue Jiang, Dingdong WANG, Hanting Wang, Hai Huang, Zhou Zhao

14:55 - 15:06

FlashAudio: Rectified Flow for Fast and High-Fidelity Text-to-Audio Generation
Huadai Liu, Jialei Wang, Rongjie Huang, Yang Liu, Heng Lu, Wei Xue, Zhou Zhao
Huadai Liu, Jialei Wang, Rongjie Huang, Yang Liu, Heng Lu, Wei Xue, Zhou Zhao

15:06 - 15:17

Integrating Audio, Visual, and Semantic Information for Enhanced Multimodal Speaker Diarization on Multi-party Conversation
Luyao Cheng, Hui Wang, Chong Deng, Siqi Zheng, Yafeng Chen, Rongjie Huang, Qinglin Zhang, Qian Chen, Xihao Li, Wen Wang
Luyao Cheng, Hui Wang, Chong Deng, Siqi Zheng, Yafeng Chen, Rongjie Huang, Qinglin Zhang, Qian Chen, Xihao Li, Wen Wang

15:17 - 15:28

Analyzing and Mitigating Inconsistency in Discrete Speech Tokens for Neural Codec Language Models
Wenrui Liu, Zhifang Guo, Jin Xu, Yuanjun Lv, Yunfei Chu, Zemin Liu, Junyang Lin
Wenrui Liu, Zhifang Guo, Jin Xu, Yuanjun Lv, Yunfei Chu, Zemin Liu, Junyang Lin

Session 9: Summarization (Tuesday 29 July, 14:00-15:30, Hall N.2)

14:11 - 14:22

uMedSum: A Unified Framework for Advancing Medical Abstractive Summarization
Aishik Nagar, Yutong Liu, Andy T. Liu, Viktor Schlegel, Vijay Prakash Dwivedi, Arun-Kumar Kaliya-Perumal, GUNA PRATHEEP KALANCHIAM, Yili Tang, Robby T. Tan
Aishik Nagar, Yutong Liu, Andy T. Liu, Viktor Schlegel, Vijay Prakash Dwivedi, Arun-Kumar Kaliya-Perumal, GUNA PRATHEEP KALANCHIAM, Yili Tang, Robby T. Tan

14:33 - 14:44

An Empirical Study of Many-to-Many Summarization with Large Language Models
Jiaan Wang, Fandong Meng, Zengkui Sun, Yunlong Liang, Yuxuan Cao, Jiarong Xu, HAOXIANG SHI, Jie Zhou
Jiaan Wang, Fandong Meng, Zengkui Sun, Yunlong Liang, Yuxuan Cao, Jiarong Xu, HAOXIANG SHI, Jie Zhou

14:55 - 15:06

Tree-of-Debate: Multi-Persona Debate Trees Elicit Critical Thinking for Scientific Comparative Analysis
Priyanka Kargupta, Ishika Agarwal, Tal August, Jiawei Han
Priyanka Kargupta, Ishika Agarwal, Tal August, Jiawei Han

Session 9: TACL Journal (Tuesday 29 July, 14:00-15:30, Hall B)

15:06 - 15:17

Benchmarking Uncertainty Quantification Methods for Large Language Models with LM-Polygraph

Roman Vashurin Ekaterina Fadeeva Artem Vazhentsev Lyudmila Rvanova Akim Tsvigun Daniil Vasilev Rui Xing Abdelrahman "Boda" Sadallah Kirill Grishchenkov Sergey Petrakov Alexander Panchenko Timothy Baldwin Preslav Nakov Maxim Panov Artem Shelmanov

The rapid proliferation of large language models (LLMs) has stimulated researchers to seek effective and efficient approaches to deal with LLM hallucinations and low-quality outputs. Uncertainty quantification (UQ) is a key element of machine learning applications in dealing with such challenges. However, research to date on UQ for LLMs has been fragmented in terms of techniques and evaluation methodologies. In this work, we address this issue by introducing a novel benchmark that implements a collection of state-of-the-art UQ baselines and offers an environment for controllable and consistent evaluation of novel UQ techniques over various text generation tasks. Our benchmark also supports the assessment of confidence normalization methods in terms of their ability to provide interpretable scores. Using our benchmark, we conduct a large-scale empirical investigation of UQ and normalization techniques across eleven tasks, identifying the most effective approaches.

15:17 - 15:28

BroadLLM: Wider and Deeper LLM Networks are Fairer LLM Evaluators

Xinghua Zhang Bowen Yu Haiyang Yu Yangyu Lv Tingwen Liu Fei Huang Hongbo Xua Yongbin Li

Measuring the quality of responses generated by LLMs is a challenging task, particularly when it comes to evaluating whether the response is aligned with human preference. A novel approach involves using the LLM itself to make evaluation and stabilizing the results through multiple independent evaluations, similar to a single-layer narrow LLM network. This network consists of a fixed number of nodes, with each node being the same LLM. This paper draws upon the extensive research to explore whether wider and deeper networks can lead to fairer evaluations. Specifically, we first adaptively generate diverse node roles for each evaluation sample. Each perspective corresponds to the role of a specific LLM node in the first layer. In subsequent layers, we integrate the locally learned evaluation information from all nodes in the previous layer to obtain a more comprehensive evaluation result. Interestingly, this network design resembles the process of academic paper reviewing, where reviewers rate independently and then consider others' opinions to reach a final decision. To validate the effectiveness, we construct the largest and most diverse evaluation benchmark LLMEval2 for LLM evaluators. Experimental results demonstrate that a wider network with 2 layers performs the best (named as BroadLLM).1 It also boosts Chinese LLM evaluation by 4.6×, saving 60% costs with 93% human agreement.

15:17 - 15:28

Contextualized Evaluations: Taking the Guesswork Out of Language Model Evaluations

Chaitanya Malaviya Joseph Chee Chang Dan Roth Mohit Iyyer Mark Yatskar Kyle Loa

Language model users often issue queries that lack specification, where the context under which a query was issued---such as the user's identity, the query's intent, and the criteria for a response to be useful---is not explicit. For instance, a good response to a subjective query like “What book should I read next?” would depend on the user’s preferences, and a good response to an open-ended query like “How do antibiotics work against bacteria?” would depend on the user's expertise. This makes evaluation of responses to such queries an ill-posed task, as evaluators must make uninformed judgements about the quality of a response. To remedy this, we present contextualized evaluations, a protocol that synthetically constructs context surrounding an underspecified query and provides it during evaluation. We find that the presence of context can 1) change conclusions drawn from evaluation, even flipping win rates between model pairs, 2) nudge evaluators to make fewer judgments based on surface-level criteria of responses, like style, and 3) provide new insights about model behavior in different contexts. Specifically, we discover an implicit bias towards WEIRD contexts in default responses and find that models are not equally sensitive to following different contexts.

Session 9: TACL Journal (Tuesday 29 July, 14:00-15:30, Room 1.85)

15:17 - 15:28

Culturally Aware and Adapted NLP: A Taxonomy and a Survey of the State of the Art

Chen Cecilia Liu Iryna Gurevych Anna Korhonen

The surge of interest in culture in NLP has inspired much recent research, but a shared understanding of "culture" remains unclear, making it difficult to evaluate progress in this emerging area. Drawing on prior research in NLP and related fields, we propose a fine-grained taxonomy of elements in culture that can provide a systematic framework for analyzing and understanding research progress. Using the taxonomy, we survey existing resources and methods for culturally aware and adapted NLP, providing an overview of the state of the art and the research gaps that still need to be filled.

Poster Presentation

Session 7: Computational Social Science and Cultural Analytics (Tuesday 29 July, 10:30-12:00, Hall 4/5)

Do Large Language Models Represent the People? A Systematic Literature Review on the Demographic Representativeness of Large Language Models

Indira Sen, Marlene Lutz, Elisa Rogers, David Garcia, Markus Strohmaier

Many applications of Large Language Models (LLMs) require them to either simulate people or offer personalized functionality, making the demographic representativeness of LLMs crucial for equitable utility. At the same time, we know little about the extent to which these models actually reflect the demographic attributes and behaviors of certain groups or populations in valid ways. Empirical research is divided, with conflicting findings on whether LLMs can accurately represent human (sub)populations. To shed light on this debate, we review 211 papers on the demographic representativeness of LLMs. We find that while 29% of the studies report positive conclusions, 30% of these do not evaluate LLMs across multiple demographic categories or within demographic subcategories. Another 35-47% fail to specify these subcategories altogether for gender and race. Out of the papers that do report subcategories, a majority do not include results on marginalized groups in their study. This suggests an inflated perception of LLM representativeness in the broader community. We recommend more precise evaluation methods and comprehensive documentation to ensure the responsible use of LLMs for social applications.

Session 7: Ethics, Bias, and Fairness (Tuesday 29 July, 10:30-12:00, Hall 4/5)

Low-Entropy Watermark Detection via Bayes' Rule Derived Detector

Beining Huang, Du Su, Fei Sun, Qi Cao, Huawei Shen, Xueqi Cheng

Text watermarking, which modify tokens to embed watermark, has proven effective in detecting machine-generated texts. Yet its application to low-entropy texts like code and mathematics presents significant challenges. A fair number of tokens in these texts are hardly modifiable without changing the intended meaning, causing statistical measures to falsely indicate the absence of a watermark. Existing research addresses this issue by rely mainly on a limited number of high-entropy tokens, which are considered flexible for modification, and accurately reflecting watermarks. However, their detection accuracy remains suboptimal, as they neglect strong watermark evidences embedded in low entropy tokens modified through watermarking. To overcome this limitation, we introduce Bayes' Rule derived Watermark Detector (BRWD), which exploit watermark information from every token, by leveraging the posterior probability of watermark's presence. We theoretically prove the optimality of our method in terms of detection accuracy, and demonstrate its superiority across various datasets, models, and watermark injection strategies. Notably, our method achieves up to 50% and 70% relative improvements in detection accuracy over the best baselines in code generation and math problem-solving tasks, respectively. Our code is available at https://github.com/cczslp/BRWD.

Session 7: Generation (Tuesday 29 July, 10:30-12:00, Hall 4/5)

Judge as A Judge: Improving the Evaluation of Retrieval-Augmented Generation through the Judge-Consistency of Large Language Models

Shuliang Liu, Xinze Li, Zhenghao Liu, Yukun Yan, Cheng Yang, Zheni Zeng, Zhiyuan Liu, Maosong Sun, Ge Yu

Retrieval-Augmented Generation (RAG) has proven its effectiveness in alleviating hallucinations for Large Language Models (LLMs). However, existing automated evaluation metrics cannot fairly evaluate the outputs generated by RAG models during training and evaluation. LLM-based judgment models provide the potential to produce high-quality judgments, but they are highly sensitive to evaluation prompts, leading to inconsistencies when judging the output of RAG models. This paper introduces the Judge-Consistency (ConsJudge) method, which aims to enhance LLMs to generate more accurate evaluations for RAG models. Specifically, ConsJudge prompts LLMs to generate different judgments based on various combinations of judgment dimensions, utilize the judge-consistency to evaluate these judgments and select the accepted and rejected judgments for DPO training. Our experiments show that ConsJudge can effectively provide more accurate judgments for optimizing RAG models across various RAG models and datasets. Further analysis reveals that judgments generated by ConsJudge have a high agreement with the superior LLM. All codes will be released via GitHub.

Session 7: Information Extraction (Tuesday 29 July, 10:30-12:00, Hall 4/5)

Generating Diverse Training Samples for Relation Extraction with Large Language Models

Zexuan Li, Hongliang Dai, Piji Li

Using Large Language Models (LLMs) to generate training data can potentially be a preferable way to improve zero or few-shot NLP tasks. However, many problems remain to be investigated for this direction. For the task of Relation Extraction (RE), we find that samples generated by directly prompting LLMs may easily have high structural similarities with each other. They tend to use a limited variety of phrasing while expressing the relation between a pair of entities. Therefore, in this paper, we study how to effectively improve the diversity of the training samples generated with LLMs for RE, while also maintaining their correctness. We first try to make the LLMs produce dissimilar samples by directly giving instructions in In-Context Learning (ICL) prompts. Then, we propose an approach to fine-tune LLMs for diversity training sample generation through Direct Preference Optimization (DPO). Our experiments on commonly used RE datasets show that both attempts can improve the quality of the generated training data. We also find that comparing with directly performing RE with an LLM, training a non-LLM RE model with its generated samples may lead to better performance.

Open-World Attribute Mining for E-Commerce Products with Multimodal Self-Correction Instruction Tuning

Jiaqi Li, Yanming Li, Xiaoli Shen, Chuanyi Zhang, Guilin Qi, Sheng Bi

In e-commerce, effective product Attribute Mining (AM) is essential for improving product features and aiding consumer decisions. However, current AM methods often focus on extracting attributes from unimodal text, underutilizing multimodal data. In this paper, we propose a novel framework called Multimodal Self-Correction Instruction Tuning (MSIT) to mine new potential attributes from both images and text with Multimodal Large Language Models. The tuning process involves two datasets: Attribute Generation Tuning Data (AGTD) and Chain-of-Thought Tuning Data (CTTD). AGTD is constructed utilizing in-context learning with a small set of seed attributes, aiding the MLLM in accurately extracting attribute-value pairs from multimodal information. To introduce explicit reasoning and improve the extraction in accuracy, we construct CTTD, which incorporates a structured 5-step reasoning process for self-correction. Finally, we employ a 3-stage inference process to filter out redundant attributes and sequentially validate each generated attribute. Comprehensive experimental results on two datasets show that MSIT outperforms state-of-the-art methods. We will release our code and data in the near future.

Towards a More Generalized Approach in Open Relation Extraction

Qing Wang, Yuepei Li, Qiao Qiao, Kang Zhou, Qi Li

Open Relation Extraction (OpenRE) seeks to identify and extract novel relational facts between named entities from unlabeled data without pre-defined relation schemas. Traditional OpenRE methods typically assume that the unlabeled data consists solely of novel relations or is pre-divided into known and novel instances. However, in real-world scenarios, novel relations are arbitrarily distributed. In this paper, we propose a generalized OpenRE setting that considers unlabeled data as a mixture of both known and novel instances. To address this, we propose MixORE, a two-phase framework that integrates relation classification and clustering to jointly learn known and novel relations. Experiments on three benchmark datasets demonstrate that MixORE consistently outperforms competitive baselines in known relation classification and novel relation clustering. Our findings contribute to the advancement of generalized OpenRE research and real-world applications.

Towards Effective Extraction and Evaluation of Factual Claims

Dasha Metropolitansky, Jonathan Larson

A common strategy for fact-checking long-form content generated by Large Language Models (LLMs) is extracting simple claims that can be verified independently. Since inaccurate or incomplete claims compromise fact-checking results, ensuring claim quality is critical. However, the lack of a standardized evaluation framework impedes assessment and comparison of claim extraction methods. To address this gap, we propose a framework for evaluating claim extraction in the context of fact-checking along with automated, scalable, and replicable methods for applying this framework, including novel approaches for measuring coverage and decontextualization. We also introduce Claimify, an LLM-based claim extraction method, and demonstrate that it outperforms existing methods under our evaluation framework. A key feature of Claimify is its ability to handle ambiguity and extract claims only when there is high confidence in the correct interpretation of the source text.

SEOE: A Scalable and Reliable Semantic Evaluation Framework for Open Domain Event Detection

Yi-Fan Lu, Xian-Ling Mao, Tian Lan, Tong Zhang, Yu-Shi Zhu, Heyan Huang

Automatic evaluation for Open Domain Event Detection (ODED) is a highly challenging task, because ODED is characterized by a vast diversity of un-constrained output labels from various domains. Nearly all existing evaluation methods for ODED usually first construct evaluation benchmarks with limited labels and domain coverage, and then evaluate ODED methods using metrics based on token-level label matching rules. However, this kind of evaluation framework faces two issues: (1) The limited evaluation benchmarks lack representatives of the real world, making it difficult to accurately reflect the performance of various ODED methods in real-world scenarios; (2) Evaluation metrics based on token-level matching rules fail to capture semantic similarity between predictions and golden labels. To address these two problems above, we propose a scalable and reliable Semantic-level Evaluation framework for Open domain Event detection (SEOE) by constructing a more representative evaluation benchmark and introducing a semantic evaluation metric. Specifically, our proposed framework first constructs a scalable evaluation benchmark that currently includes 564 event types covering 7 major domains, with a cost-effective supplementary annotation strategy to ensure the benchmark's representativeness. The strategy also allows for the supplement of new event types and domains in the future. Then, the proposed SEOE leverages large language models (LLMs) as automatic evaluation agents to compute a semantic F1-score, incorporating fine-grained definitions of semantically similar labels to enhance the reliability of the evaluation. Extensive experiments validate the representatives of the benchmark and the reliability of the semantic evaluation metric. Existing ODED methods are thoroughly evaluated, and the error patterns of predictions are analyzed, revealing several insightful findings.

Selecting and Merging: Towards Adaptable and Scalable Named Entity Recognition with Large Language Models

Zhuojun Ding, Wei Wei, Chenghao Fan

Supervised fine-tuning (SFT) is widely used to align large language models (LLMs) with information extraction (IE) tasks, such as named entity recognition (NER). However, annotating such fine-grained labels and training domain-specific models is costly. Existing works typically train a unified model across multiple domains, but such approaches lack adaptation and scalability since not all training data benefits target domains and scaling trained models remains challenging. We propose the SaM framework, which dynamically Selects and Merges expert models at inference time. Specifically, for a target domain, we select domain-specific experts pre-trained on existing domains based on (i) domain similarity to the target domain and (ii) performance on sampled instances, respectively. The experts are then merged to create task-specific models optimized for the target domain. By dynamically merging experts beneficial to target domains, we improve generalization across various domains without extra training. Additionally, experts can be added or removed conveniently, leading to great scalability. Extensive experiments on multiple benchmarks demonstrate our framework’s effectiveness, which outperforms the unified model by an average of 10%. We further provide insights into potential improvements, practical experience, and extensions of our framework.

A Generative Adaptive Replay Continual Learning Model for Temporal Knowledge Graph Reasoning

Zhiyu Zhang, Wei Chen, Youfang Lin, Huaiyu Wan

Recent Continual Learning (CL)-based Temporal Knowledge Graph Reasoning (TKGR) methods focus on significantly reducing computational cost and mitigating catastrophic forgetting caused by fine-tuning models with new data. However, existing CL-based TKGR methods still face two key limitations: (1) They usually one-sidedly reorganize individual historical facts, while overlooking the historical context essential for accurately understanding the historical semantics of these facts; (2) They preserve historical knowledge by simply replaying historical facts, while ignoring the potential conflicts between historical and emerging facts. In this paper, we propose a \textbf{D}eep \textbf{G}enerative \textbf{A}daptive \textbf{R}eplay (DGAR) method, which can generate and adaptively replay historical entity distribution representations from the whole historical context. To address the first challenge, historical context prompts as sampling units are built to preserve the whole historical context information. To overcome the second challenge, a pre-trained diffusion model is adopted to generate the historical distribution. During the generation process, the common features between the historical and current distributions are enhanced under the guidance of the TKGR model. In addition, a layer-by-layer adaptive replay mechanism is designed to effectively integrate historical and current distributions. Experimental results demonstrate that DGAR significantly outperforms baselines in reasoning and mitigating forgetting.

RSCF: Relation-Semantics Consistent Filter for Entity Embedding of Knowledge Graph

Junsik Kim, Jinwook Park, Kangil Kim

In knowledge graph embedding, leveraging relation specific entity transformation has markedly enhanced performance. However, the consistency of embedding differences before and after transformation remains unaddressed, risking the loss of valuable inductive bias inherent in the embeddings. This inconsistency stems from two problems. First, transformation representations are specified for relations in a disconnected manner, allowing dissimilar transformations and corresponding entity embeddings for similar relations. Second, a generalized plug-in approach as a SFBR (Semantic Filter Based on Relations) disrupts this consistency through excessive concentration of entity embeddings under entity-based regularization, generating indistinguishable score distributions among relations. In this paper, we introduce a plug-in KGE method, Relation-Semantics Consistent Filter (RSCF). Its entity transformation has three features for enhancing semantic consistency: 1) shared affine transformation of relation embeddings across all relations, 2) rooted entity transformation that adds an entity embedding to its change represented by the transformed vector, and 3) normalization of the change to prevent scale reduction. To amplify the advantages of consistency that preserve semantics on embeddings, RSCF adds relation transformation and prediction modules for enhancing the semantics. In knowledge graph completion tasks with distance-based and tensor decomposition models, RSCF significantly outperforms state-of-the-art KGE methods, showing robustness across all relations and their frequencies.

Removal of Hallucination on Hallucination: Debate-Augmented RAG

Wentao Hu, Wengyu Zhang, Yiyang Jiang, Chen Jason Zhang, Xiaoyong Wei, Li Qing

Retrieval-Augmented Generation (RAG) enhances factual accuracy by integrating external knowledge, yet it introduces a critical issue: erroneous or biased retrieval can mislead generation, compounding hallucinations, a phenomenon we term Hallucination on Hallucination. To address this, we propose Debate-Augmented RAG (DRAG), a training-free framework that integrates Multi-Agent Debate (MAD) mechanisms into both retrieval and generation stages. In retrieval, DRAG employs structured debates among proponents, opponents, and judges to refine retrieval quality and ensure factual reliability. In generation, DRAG introduces asymmetric information roles and adversarial debates, enhancing reasoning robustness and mitigating factual inconsistencies. Evaluations across multiple tasks demonstrate that DRAG improves retrieval reliability, reduces RAG-induced hallucinations, and significantly enhances overall factual accuracy. Our code is available at https://github.com/Huenao/Debate-Augmented-RAG.

TST: A Schema-Based Top-Down and Dynamic-Aware Agent of Text-to-Table Tasks

Peiwen Jiang, Haitong Jiang, Ruhui Ma, Yvonne Jie Chen, Jinhua Cheng

As a bridge between natural texts and information systems like structured storage, statistical analysis, retrieving, and recommendation, the text-to-table task has received widespread attention recently. Existing researches have gone through a paradigm shift from traditional bottom-up IE (Information Extraction) to top-down LLMs-based question answering with RAG (Retrieval-Augmented Generation). Furthermore, these methods mainly adopt end-to-end models or use multi-stage pipelines to extract text content based on static table structures. However, they neglect to deal with precise inner-document evidence extraction and dynamic information such as multiple entities and events, which can not be defined in static table head format and are very common in natural texts. To address this issue, we propose a two-stage dynamic content extraction agent framework called TST (Text-Schema-Table), which uses type recognition methods to extract context evidences with the conduction of domain schema sequentially. Based on the evidence, firstly we quantify the total instances of each dynamic object and then extract them with ordered numerical prompts. Through extensive comparisons with existing methods across different datasets, our extraction framework exhibits state-of-the-art (SOTA) performance. Our codes are available at .

Tree-KG: An Expandable KG Construction Framework for Knowledge-intensive Domains

Songjie Niu, Kaisen Yang, Rui Zhao, Yichao Liu, Zonglin Li, Hongning Wang, Wenguang Chen

In knowledge-intensive domains like scientific research, effective decisions rely on organizing and retrieving intricate data. Knowledge graphs (KGs) help by structuring entities, relations, and contextual dependencies, but building KGs in such domains is challenging due to inherent complexity, manual effort, and rapid evolution. Inspired by how humans organize knowledge hierarchically, we propose Tree-KG, an expandable framework that combines structured domain texts with advanced semantic techniques. First, Tree-KG builds a tree-like graph from textbook structures using large language models (LLMs) and domain-specific entities, creating an \emph{explicit KG}. Then, through iterative expansion with flexible, predefined operators, it uncovers \emph{hidden KG} while preserving semantic coherence. Experiments demonstrate that Tree-KG consistently surpasses competing methods, achieving the highest F1 scores (12–16\% above the second-best), with notable performance (F1 0.81) on the Text-Annotated dataset, highlighting its effectiveness in extracting high-quality information from source texts. Additionally, Tree-KG provides superior structural alignment, domain-specific extraction, and cost-efficiency, delivering robust results with reduced token usage and adaptable, resource-conscious deployment. \footnote{}

Employing Discourse Coherence Enhancement to Improve Cross-Document Event and Entity Coreference Resolution

Xinyu Chen, Peifeng Li, Qiaoming Zhu

Cross-Document Coreference Resolution (CDCR) aims to identify and group together mentions of a specific event or entity that occur across multiple documents. In contrast to the within-document tasks, in which event and entity mentions are linked by rich and coherent contexts, cross-document mentions lack such critical contexts, which presents a significant challenge in establishing connections among them. To address this issue, we introduce a novel task Cross-Document Discourse Coherence Enhancement (CD-DCE) to enhance the discourse coherence between two cross-document event or entity mentions. Specifically, CD-DCE first selects coherent texts and then adds them between two cross-document mentions to form a new coherent document. Subsequently, the coherent text is employed to represent the event or entity mentions and to resolve any coreferent mentions. Experimental results on the three popular datasets demonstrate that our proposed method outperforms several state-of-the-art baselines.

Temporal Relation Extraction in Clinical Texts: A Span-based Graph Transformer Approach

Rochana Chaturvedi, Peyman Baghershahi, Sourav Medya, Barbara Di Eugenio

Temporal information extraction from unstructured text is essential for contextualizing events and deriving actionable insights, particularly in the medical domain. We address the task of extracting clinical events and their temporal relations using the well-studied I2B2 2012 Temporal Relations Challenge corpus. This task is inherently challenging due to complex clinical language, long documents, and sparse annotations. We introduce GraphTREx, a novel method integrating span-based entity-relation extraction, clinical large pre-trained language models (LPLMs), and Heterogeneous Graph Transformers (HGT) to capture local and global dependencies. Our HGT component facilitates information propagation across the document through innovative global landmarks that bridge distant entities and improves the state-of-the-art with 5.5% improvement in the tempeval F1 score over the previous best and up to 8.9% improvement on long-range relations, which presents a formidable challenge. We further demonstrate generalizability by establishing a strong baseline on the E3C corpus. Not only does this work advance temporal information extraction, but also lays the groundwork for improved diagnostic and prognostic models through enhanced temporal reasoning.

Query-driven Document-level Scientific Evidence Extraction from Biomedical Studies

Massimiliano Pronesti, Joao H Bettencourt-Silva, Paul Flanagan, Alessandra Pascale, Oisín Redmond, Anya Belz, Yufang Hou

Extracting scientific evidence from biomedical studies for clinical research questions (e.g., Does stem cell transplantation improve quality of life in patients with medically refractory Crohn's disease compared to placebo?) is a crucial step in synthesising biomedical evidence. In this paper, we focus on the task of document-level scientific evidence extraction for clinical questions with conflicting evidence. To support this task, we create a dataset called CochraneForest leveraging forest plots from Cochrane systematic reviews. It comprises 202 annotated forest plots, associated clinical research questions, full texts of studies, and study-specific conclusions. Building on CochraneForest, we propose URCA (Uniform Retrieval Clustered Augmentation), a retrieval-augmented generation framework designed to tackle the unique challenges of evidence extraction. Our experiments show that URCA outperforms the best existing methods by up to 10.3% in F1 score on this task. However, the results also underscore the complexity of CochraneForest, establishing it as a challenging testbed for advancing automated evidence synthesis systems.

Towards Robust Universal Information Extraction: Dataset, Evaluation, and Solution

Jizhao Zhu, Akang Shi, Zixuan Li, Long Bai, Xiaolong Jin, Jiafeng Guo, Xueqi Cheng

In this paper, we aim to enhance the robustness of Universal Information Extraction (UIE) by introducing a new benchmark dataset, a comprehensive evaluation, and a feasible solution. Existing robust benchmark datasets have two key limitations: 1) They generate only a limited range of perturbations for a single Information Extraction (IE) task, which fails to evaluate the robustness of UIE models effectively; 2) They rely on small models or handcrafted rules to generate perturbations, often resulting in unnatural adversarial examples. Considering the powerful generation capabilities of Large Language Models (LLMs), we introduce a new benchmark dataset for Robust UIE, called RUIE-Bench, which utilizes LLMs to generate more diverse and realistic perturbations across different IE tasks. Based on this dataset, we comprehensively evaluate existing UIE models and reveal that both LLM-based models and other models suffer from significant performance drops. To improve robustness and reduce training costs, we propose a data-augmentation solution that dynamically selects hard samples for iterative training based on the model's inference loss. Experimental results show that training with only \textbf{15}\% of the data leads to an average \textbf{8.1}\% relative performance improvement across three IE tasks. Our code and dataset are available at: https://github.com/ICT-GoKnow/RobustUIE.

GETReason: Enhancing Image Context Extraction through Hierarchical Multi-Agent Reasoning

Shikhhar Siingh, Abhinav Rawat, Chitta Baral, Vivek Gupta

Publicly significant images from events carry valuable contextual information with applications in domains such as journalism and education. However, existing methodologies often struggle to accurately extract this contextual relevance from images. To address this challenge, we introduce GETREASON (Geospatial Event Temporal Reasoning), a framework designed to go beyond surfacelevel image descriptions and infer deeper contextual meaning. We hypothesize that extracting global event, temporal, and geospatial information from an image enables a more accurate understanding of its contextual significance. We also introduce a new metric GREAT (Geospatial, Reasoning and Event Accuracy with Temporal alignment) for a reasoning capturing evaluation. Our layered multi-agentic approach, evaluated using a reasoning-weighted metric, demonstrates that meaningful information can be inferred from images, allowing them to be effectively linked to their corresponding events and broader contextual background.

TaxoAdapt: Aligning LLM-Based Multidimensional Taxonomy Construction to Evolving Research Corpora

Priyanka Kargupta, Nan Zhang, Yunyi Zhang, Rui Zhang, Prasenjit Mitra, Jiawei Han

The rapid evolution of scientific fields introduces challenges in organizing and retrieving scientific literature. While expert-curated taxonomies have traditionally addressed this need, the process is time-consuming and expensive. Furthermore, recent automatic taxonomy construction methods either (1) over-rely on a specific corpus, sacrificing generalizability, or (2) depend heavily on the general knowledge of large language models (LLMs) contained within their pre-training datasets, often overlooking the dynamic nature of evolving scientific domains. Additionally, these approaches fail to account for the multi-faceted nature of scientific literature, where a single research paper may contribute to multiple dimensions (e.g., methodology, new tasks, evaluation metrics, benchmarks). To address these gaps, we propose TaxoAdapt, a framework that dynamically adapts an LLM-generated taxonomy to a given corpus across multiple dimensions. TaxoAdapt performs iterative hierarchical classification, expanding both the taxonomy width and depth based on corpus' topical distribution. We demonstrate its state-of-the-art performance across a diverse set of computer science conferences over the years to showcase its ability to structure and capture the evolution of scientific fields. As a multidimensional method, TaxoAdapt generates taxonomies that are 26.51% more granularity-preserving and 50.41% more coherent than the most competitive baselines judged by LLMs.

Enhancing NER by Harnessing Multiple Datasets with Conditional Variational Autoencoders

Taku Oi, Makoto Miwa

We propose a novel method to integrate a Conditional Variational Autoencoder (CVAE) into a span-based Named Entity Recognition (NER) model to model the shared and unshared information among labels in multiple datasets and ease the training on the datasets. Experimental results using multiple biomedical datasets show the effectiveness of the proposed method, achieving improved performance on the BioRED dataset.

Multimodal Coreference Resolution for Chinese Social Media Dialogues: Dataset and Benchmark Approach

Xingyu Li, Chen Gong, Guohong Fu

Multimodal coreference resolution (MCR) aims to identify mentions referring to the same entity across different modalities, such as text and visuals, and is essential for understanding multimodal content. In the era of rapidly growing multimodal content and social media, MCR is particularly crucial for interpreting user interactions and bridging text-visual references to improve communication and personalization. However, MCR research for real-world dialogues remains unexplored due to the lack of sufficient data resources. To address this gap, we introduce TikTalkCoref, the first Chinese multimodal coreference dataset for social media in real-world scenarios, derived from the popular Douyin short-video platform. This dataset pairs short videos with corresponding textual dialogues from user comments and includes manually annotated coreference clusters for both person mentions in the text and the coreferential person head regions in the corresponding video frames. We also present an effective benchmark approach for MCR, focusing on the celebrity domain, and conduct extensive experiments on our dataset, providing reliable benchmark results for this newly constructed dataset. We release the TikTalkCoref dataset to facilitate future research on MCR for real-world social media dialogues at https://github.com/lxystaruni/TikTalkCoref.

Counterfactual-Consistency Prompting for Relative Temporal Understanding in Large Language Models

Jongho Kim, seung-won hwang

Despite the advanced capabilities of large language models (LLMs), their temporal reasoning ability remains underdeveloped. Prior works have highlighted this limitation, particularly in maintaining temporal consistency when understanding event relations. For example, models often confuse mutually exclusive temporal relations like “before” and “after” between events and make inconsistent predictions. In this work, we tackle the issue of temporal inconsistency in LLMs by proposing a novel counterfactual prompting approach. Our method generates counterfactual questions and enforces collective constraints, enhancing the model’s consistency. We evaluate our method on multiple datasets, demonstrating significant improvements in event ordering for explicit and implicit events and temporal commonsense understanding, by effectively addressing temporal inconsistencies.

Session 7: Information Retrieval and Text Mining (Tuesday 29 July, 10:30-12:00, Hall 4/5)

Nested-Refinement Metamorphosis: Reflective Evolution of Prompts and Code for Efficient Algorithm Design

Shuhan Guo, Nan Yin, James Kwok, Quanming Yao

Large Language Models (LLMs) have become powerful tools for algorithm design, leveraging extensive knowledge and contextual reasoning. However, existing methods rely on iterative code generation and performance-based filtering, leading to inefficient exploration and high computational costs. Inspired by butterfly metamorphosis, where organisms undergo structured developmental transitions, Phase I allows larvae to accumulate nutrients, laying the foundation for Phase II, where the pupa transforms into a butterfly. Building on this principle, we introduce Metamorphosis on Prompts (MoP) to iteratively refine task descriptions and Metamorphosis on Algorithms (MoA) to generate more effective solutions. Their nested refinement ensures task-algorithm alignment, systematically improving both task descriptions and algorithmic solutions for more efficient algorithm design. To further enhance efficiency, we incorporate predictor-assisted code evaluation, mimicking natural selection by filtering out weak candidates early and reducing computational costs. Experimental results on TSP, MKP, and CVRP demonstrate that NeRM consistently outperforms state-of-the-art approaches in both performance and efficiency.

Session 7: Interpretability and Analysis of Models for NLP (Tuesday 29 July, 10:30-12:00, Hall 4/5)

REVS: Unlearning Sensitive Information in Language Models via Rank Editing in the Vocabulary Space

Tomer Ashuach, Martin Tutek, Yonatan Belinkov

Language models (LMs) risk inadvertently memorizing and divulging sensitive or personally identifiable information (PII) seen in training data, causing privacy concerns. Current approaches to address this issue involve costly dataset scrubbing, or model filtering through unlearning and model editing, which can be bypassed through extraction attacks. We propose REVS, a novel non-gradient-based method for unlearning sensitive information from LMs. REVS identifies and modifies a small subset of neurons relevant for constituent tokens which form sensitive information. To adequately evaluate our method on truly sensitive information, we curate three datasets: an email and URL datasets naturally memorized by the models, and a synthetic social security number dataset that we tune the models to memorize. Compared to other methods, REVS demonstrates superior performance in unlearning sensitive information and robustness to extraction attacks, while retaining underlying model integrity.

How Do Multilingual Language Models Remember Facts?

Constanza Fierro, Negar Foroutan, Desmond Elliott, Anders Søgaard

Large Language Models (LLMs) store and retrieve vast amounts of factual knowledge acquired during pre-training. Prior research has localized and identified mechanisms behind knowledge recall; however, it has only focused on English monolingual models. The question of how these mechanisms generalize to non-English languages and multilingual LLMs remains unexplored. In this paper, we address this gap by conducting a comprehensive analysis of three multilingual LLMs. First, we show that previously identified recall mechanisms in English largely apply to multilingual contexts, with nuances based on language and architecture. Next, through patching intermediate representations, we localize the role of language during recall, finding that subject enrichment is language-independent, while object extraction is language-dependent. Additionally, we discover that the last token representation acts as a Function Vector (FV), encoding both the language of the query and the content to be extracted from the subject. Furthermore, in decoder-only LLMs, FVs compose these two pieces of information in two separate stages. These insights reveal unique mechanisms in multilingual LLMs for recalling information, highlighting the need for new methodologies—such as knowledge evaluation, fact editing, and knowledge acquisition—that are specifically tailored for multilingual LLMs.

Eliciting Textual Descriptions from Representations of Continuous Prompts

Daniela Gottesman, Mor Geva, Dana Ramati

Continuous prompts, or "soft prompts", are a widely-adopted parameter-efficient tuning strategy for large language models, but are often less favorable due to their opaque nature. Prior attempts to interpret continuous prompts relied on projecting individual prompt tokens onto the vocabulary space. However, this approach is problematic as performant prompts can yield arbitrary or contradictory text, and it individually interprets each prompt token. In this work, we propose a new approach to interpret continuous prompts that elicits textual descriptions from their representations during model inference. Using a Patchscopes variant (Ghandeharioun et al., 2024) called InSPEcT over various tasks, we show our method often yields accurate task descriptions which become more faithful as task performance increases. Moreover, an elaborated version of InSPEcT reveals biased features in continuous prompts, whose presence correlates with biased model predictions. Providing an effective interpretability solution, InSPEcT can be leveraged to debug unwanted properties in continuous prompts and inform developers on ways to mitigate them.

Axiomatic Analysis of Uncertainty Estimation For Retrieval Augmented Generation

Heydar Soudani, Evangelos Kanoulas, Faegheh Hasibi

Large Language Models (LLMs) are valued for their strong performance across various tasks, but they also produce inaccurate or misleading outputs. Uncertainty Estimation (UE) quantifies the model's confidence and helps users assess response reliability. However, existing UE methods have not been thoroughly examined in scenarios like Retrieval-Augmented Generation (RAG), where the input prompt includes non-parametric knowledge. This paper shows that current UE methods cannot reliably assess correctness in the RAG setting. We further propose an axiomatic framework to identify deficiencies in existing methods and guide the development of improved approaches. Our framework introduces five constraints that an effective UE method should meet after incorporating retrieved documents into the LLM's prompt. Experimental results reveal that no existing UE method fully satisfies all the axioms, explaining their suboptimal performance in RAG. We further introduce a simple yet effective calibration function based on our framework, which not only satisfies more axioms than baseline methods but also improves the correlation between uncertainty estimates and correctness.

On Entity Identification in Language Models

Masaki Sakata, Sho Yokoi, Benjamin Heinzerling, Takumi Ito, Kentaro Inui

This study investigates the condition under which Transformer-based language models (LMs) identify and distinguish entities, such as individuals and organizations, with their internal representations, focusing on the many-to-many correspondence between entities and their mentions. We first formulate two problems, ambiguity and variability of entity mentions, and propose a framework analogous to clustering quality metrics to measure how well the models resolve these issues. Our experiments examine four Transformer-based autoregressive models, showing that they effectively identify and distinguish entities with metrics analogous to precision and recall exceeding 0.8. Further analysis reveals that entity-related information is compactly represented in a low-dimensional linear subspace at early LM layers and provides potential interpretations of these findings through the lens of isomorphism between LM representations and entity-centric knowledge structures in the real world.

The anatomy of evidence: An investigation into explainable ICD coding

Katharina Beckh, Felix Studeny, Sujan Sai Gannamaneni, Dario Antweiler, Stefan Rueping

Automatic medical coding has the potential to ease documentation and billing processes. For this task, transparency plays an important role for medical coders and regulatory bodies, which can be achieved using explainability methods. However, the evaluation of these approaches has been mostly limited to short text and binary settings due to a scarcity of annotated data. Recent efforts by Cheng et al. (2023) have introduced the MDACE dataset, which provides a valuable resource containing code evidence in clinical records. In this work, we conduct an in-depth analysis of the MDACE dataset and perform plausibility evaluation of current explainable ICD coding systems from an applied perspective. With this, we contribute to a deeper understanding of automatic ICD coding and evidence extraction. Our findings reveal that ground truth evidence aligns with code descriptions to a certain degree. An investigation into state-of-the-art approaches shows a high overlap with ground truth evidence. We propose introduce match measures and highlight success and failure cases. Based on our findings, we provide recommendations for developing and evaluating explainable ICD coding systems.

Word Form Matters: LLMs' Semantic Reconstruction under Typoglycemia

Chenxi Wang, Tianle Gu, zhongyu wei, Lang Gao, Zirui Song, Xiuying Chen

Human readers can efficiently comprehend scrambled words, a phenomenon known as \textit{Typoglycemia}, primarily by relying on word form; if word form alone is insufficient, they further utilize contextual cues for interpretation. While advanced large language models (LLMs) exhibit similar abilities, the underlying mechanisms remain unclear. To investigate this, we conduct controlled experiments to analyze the roles of word form and contextual information in semantic reconstruction and examine LLM attention patterns. Specifically, we first propose \textit{SemRecScore}, a reliable metric to quantify the degree of semantic reconstruction, and validate its effectiveness. Using this metric, we study how word form and contextual information influence LLMs' semantic reconstruction ability, identifying \textit{word form as the core factor} in this process. Furthermore, we analyze \textit{how} LLMs utilize word form and find that they rely on specialized attention heads to extract and process word form information, with this mechanism remaining stable across varying levels of word scrambling. This distinction between LLMs' fixed attention patterns primarily focused on word form and human readers' adaptive strategy in balancing word form and contextual information provides insights into enhancing LLM performance by incorporating human-like, context-aware mechanisms. Code is available on: https://anonymous.4open.science/r/typoglycemia-E0A9/.

SeqMMR: Sequential Model Merging and LLM Routing for Enhanced Batched Sequential Knowledge Editing

Shanbao Qiao, Xuebing Liu, Akshat Gupta, Seung-Hoon Na

Model knowledge editing enables the efficient correction of erroneous information and the continuous updating of outdated knowledge within language models. While existing research has demonstrated strong performance in single-instance or few-instance sequential editing and one-time massive editing scenarios, the batched sequential editing paradigm remains a significant challenge. The primary issue lies in the model's tendency to gradually forget previously edited knowledge and become increasingly unstable after multiple iterations of batched editing. To address these challenges, we propose **SeqMMR**, an enhanced framework for batched sequential knowledge editing that leverages **Seq**uential **M**odel **M**erging and a model **R**outer. Our approach iteratively merges parameters from current batch-edited models with those of their predecessors, ensuring that newly emerging knowledge is integrated while mitigating the forgetting of previously edited knowledge. Furthermore, the model router directs queries unrelated to the edited knowledge to an unedited model backup, preventing unintended alterations in model predictions. Extensive experiments across various datasets demonstrate that our approach effectively mitigates knowledge forgetting, improves performance across all previous batches, and better preserves the model's general capabilities.

FADE: Why Bad Descriptions Happen to Good Features

Bruno Puri, Aakriti Jain, Elena Golimblevskaia, Patrick Kahardipraja, Thomas Wiegand, Wojciech Samek, Sebastian Lapuschkin

Recent advances in mechanistic interpretability have highlighted the potential of automating interpretability pipelines in analyzing the latent representations within LLMs. While they may enhance our understanding of internal mechanisms, the field lacks standardized evaluation methods for assessing the validity of discovered features. We attempt to bridge this gap by introducing FADE: Feature Alignment to Description Evaluation, a scalable model-agnostic framework for evaluating feature-description alignment. FADE evaluates alignment across four key metrics - Clarity, Responsiveness, Purity, and Faithfulness - and systematically quantifies the causes for the misalignment of feature and their description. We apply FADE to analyze existing open-source feature descriptions, and assess key components of automated interpretability pipelines, aiming to enhance descriptions quality. Our findings highlight fundamental challenges in generating feature descriptions, particularly for SAEs as compared to MLP neurons, providing insights into the limitations and future directions of automated interpretability.

Metaphor and Large Language Models: When Surface Features Matter More than Deep Understanding

Elisa Sanchez-Bayona, Rodrigo Agerri

This paper presents a comprehensive evaluation of the capabilities of Large Language Models (LLMs) in metaphor interpretation across multiple datasets, tasks, and prompt configurations. Although metaphor processing has gained significant attention in Natural Language Processing (NLP), previous research has been limited to single-dataset evaluations and specific task settings, often using artificially constructed data through lexical replacement. We address these limitations by conducting extensive experiments using diverse publicly available datasets with inference and metaphor annotations, focusing on Natural Language Inference (NLI) and Question Answering (QA) tasks. The results indicate that LLMs' performance is more influenced by features like lexical overlap and sentence length than by metaphorical content, demonstrating that any alleged emergent abilities of LLMs to understand metaphorical language are the result of a combination of surface-level features, in-context learning, and linguistic knowledge. This work provides critical insights into the current capabilities and limitations of LLMs in processing figurative language, highlighting the need for more realistic evaluation frameworks in metaphor interpretation tasks. Data and code publicly available: https://anonymous.4open.science/r/metaphorLLM-047E/

Fact Recall, Heuristics or Pure Guesswork? Precise Interpretations of Language Models for Fact Completion

Denitsa Saynova, Lovisa Hagström, Moa Johansson, Richard Johansson, Marco Kuhlmann

Language models (LMs) can make a correct prediction based on many possible signals in a prompt, not all corresponding to recall of factual associations. However, current interpretations of LMs fail to take this into account. For example, given the query "Astrid Lindgren was born in" with the corresponding completion "Sweden", no difference is made between whether the prediction was based on knowing where the author was born or assuming that a person with a Swedish-sounding name was born in Sweden. In this paper, we present a model-specific recipe - PrISM - for constructing datasets with examples of four different prediction scenarios: generic language modeling, guesswork, heuristics recall and exact fact recall. We apply two popular interpretability methods to the scenarios: causal tracing (CT) and information flow analysis. We find that both yield distinct results for each scenario. Results for exact fact recall and generic language modeling scenarios confirm previous conclusions about the importance of mid-range MLP sublayers for fact recall, while results for guesswork and heuristics indicate a critical role of late last token position MLP sublayers. In summary, we contribute resources for a more extensive and granular study of fact completion in LMs, together with analyses that provide a more nuanced understanding of how LMs process fact-related queries.

How Do LLMs Acquire New Knowledge? A Knowledge Circuits Perspective on Continual Pre-Training

Yixin Ou, Yunzhi Yao, Ningyu Zhang, Hui Jin, Jiacheng Sun, Shumin Deng, Zhenguo Li, Huajun Chen

Despite exceptional capabilities in knowledge-intensive tasks, Large Language Models (LLMs) face a critical gap in understanding how they internalize new knowledge, particularly how acquired knowledge becomes structurally embedded in their neural computations. We address this issue through the lens of knowledge circuit evolution, identifying computational subgraphs that facilitate knowledge storage and processing. Our systematic analysis of circuit evolution throughout continual pre-training reveals several key findings: (1) the acquisition of new knowledge is influenced by its relevance to pre-existing knowledge; (2) the evolution of knowledge circuits exhibits a distinct phase shift from formation to optimization; (3) the evolution of knowledge circuits follows a deep-to-shallow pattern. These insights not only advance our theoretical understanding of the mechanisms of new knowledge acquisition in LLMs, but also provide potential implications for improving continual pre-training strategies to enhance model performance.

None of the Above, Less of the Right Parallel Patterns in Human and LLM Performance on Multi-Choice Questions Answering

Zhi Rui Tam, Cheng-Kuang Wu, Chieh-Yen Lin, Yun-Nung Chen

Multiple-choice exam questions with "None of the above" (NA) options have been extensively studied in educational testing, in which existing research suggests that they better assess true knowledge. However, their impact on Large Language Models (LLMs) evaluation remains underexplored. Through systematic experiments with 28 LLMs on the MMLU benchmark, we examine how NA options affect model performance and confidence calibration. Our analysis reveals that NA options, when used as the correct answer, lead to a consistent 30-50% performance drop across models regardless of scale--suggesting that LLMs lack the meta-cognitive ability to systematically evaluate and reject all given options when none are correct. This degradation shows strong domain dependence, with minimal impact on mathematical reasoning (14.6% drop) but severe effects on tasks requiring uncertainty handling like business ethics (48.1% drop). Our results highlight important implications for benchmark design and raise questions about LLMs' ability to handle uncertainty in real-world applications.

Uncertainty Unveiled: Can Exposure to More In-context Examples Mitigate Uncertainty for Large Language Models?

Yifei Wang, Yu Sheng, Linjing Li, Daniel Dajun Zeng

Recent advances in handling long sequences have unlocked new possibilities for long-context in-context learning (ICL). While existing research predominantly focuses on performance gains driven by additional in-context examples, the impact on the trustworthiness of generated responses remains underexplored. This paper addresses this gap by investigating how increased examples influence predictive uncertainty—an essential aspect in trustworthiness. We begin by systematically quantifying uncertainty across different “shot” configurations in ICL, emphasizing the role of example quantity. Through uncertainty decomposition, we introduce a novel perspective on performance enhancement, focusing on epistemic uncertainty (EU). Our results reveal that additional examples reduce total uncertainty in both simple and complex tasks by injecting task-specific knowledge, thereby diminishing EU and enhancing performance. For complex tasks, these advantages emerge only after addressing the increased noise and uncertainty associated with longer inputs. Finally, we investigate the progression of internal confidence across layers, uncovering the underlying mechanisms that drive the reduction in uncertainty.

TAGExplainer: Narrating Graph Explanations for Text-Attributed Graph Learning Models

Bo Pan, Zhen Xiong, Guanchen Wu, Zheng Zhang, Yifei Zhang, Yuntong Hu, Liang Zhao

Graph representation learning has garnered significant attention due to its broad applications in various domains, such as recommendation systems and social network analysis. Despite advancements in graph learning methods, challenges still remain in explainability when graphs are associated with semantic features. In this paper, we present GraphNarrator, the first method designed to generate natural language explanations for Graph Neural Networks. GraphNarrator employs a generative language model that maps input-output pairs to explanations reflecting the model's decision-making process. To address the lack of ground truth explanations to train the model, we propose first generating pseudo-labels that capture the model's decisions from saliency-based explanations, then using Expert Iteration to iteratively train the pseudo-label generator based on training objectives on explanation quality. The high-quality pseudo-labels are finally utilized to train an end-to-end explanation generator model. Extensive experiments are conducted to demonstrate the effectiveness of GraphNarrator in producing faithful, concise, and human-preferred natural language explanations.

Interlocking-free Selective Rationalization Through Genetic-based Learning

Federico Ruggeri, Gaetano Signorelli

A popular end-to-end architecture for selective rationalization is the select-then-predict pipeline, comprising a generator to extract highlights fed to a predictor. Such a cooperative system suffers from suboptimal equilibrium minima due to the dominance of one of the two modules, a phenomenon known as interlocking. While several contributions aimed at addressing interlocking, they only mitigate its effect, often by introducing feature-based heuristics, sampling, and ad-hoc regularizations. We present GenSPP, the first interlocking-free architecture for selective rationalization that does not require any learning overhead, as the above-mentioned. GenSPP avoids interlocking by performing disjoint training of the generator and predictor via genetic global search. Experiments on a synthetic and a real-world benchmark show that our model outperforms several state-of-the-art competitors.

HALoGEN: Fantastic LLM Hallucinations and Where to Find Them

Abhilasha Ravichander, Shrusti Ghela, David Wadden, Yejin Choi

Despite their impressive ability to generate high-quality and fluent text, generative large language models (LLMs) also produce hallucinations: statements that are misaligned with established world knowledge or provided input context. However, measuring hallucination can be challenging, as having humans verify model generations on-the-fly is both expensive and time-consuming. In this work, we release HALoGEN, a comprehensive hallucination benchmark consisting of: (1) 10,923 prompts for generative models spanning nine domains including programming, scientific attribution, and summarization, and (2) automatic high-precision verifiers for each use case that decompose LLM generations into atomic units, and verify each unit against a high-quality knowledge source. We use this framework to evaluate 150,000 generations from 14 language models, finding that even the best-performing models are riddled with hallucinations (sometimes up to 86% of generated atomic facts depending on the domain). We further define a novel error classification for LLM hallucinations based on whether they likely stem from incorrect recollection of training data (Type A errors), or incorrect knowledge in training data (Type B errors), or are fabrication (Type C errors). We hope our framework provides a foundation to enable the principled study of why generative models hallucinate, and advances the development of trustworthy large language models.

The Hidden Attention of Mamba Models

Ameen Ali Ali, Itamar Zimerman, Lior Wolf

The Mamba layer offers an efficient selective state-space model (SSM) that is highly effective in modeling multiple domains, including NLP, long-range sequence processing, and computer vision. Selective SSMs are viewed as dual models, in which one trains in parallel on the entire sequence via an IO-aware parallel scan, and deploys in an autoregressive manner. We add a third view and show that such models can be viewed as attention-driven models. This new perspective enables us to empirically and theoretically compare the underlying mechanisms to that of the attention in transformers and allows us to peer inside the inner workings of the Mamba model with explainability methods. Our code is publicly available.

Normalized AOPC: Fixing Misleading Faithfulness Metrics for Feature Attributions Explainability

Joakim Edin, Andreas Geert Motzfeldt, Casper L. Christensen, Tuukka Ruotsalo, Lars Maaløe, Maria Maistro

Deep neural network predictions are notoriously difficult to interpret. Feature attribution methods aim to explain these predictions by identifying the contribution of each input feature. Faithfulness, often evaluated using the area over the perturbation curve (AOPC), reflects feature attributions' accuracy in describing the internal mechanisms of deep neural networks. However, many studies rely on AOPC to compare faithfulness across different models, which we show can lead to false conclusions about models' faithfulness. Specifically, we find that AOPC is sensitive to variations in the model, resulting in unreliable cross-model comparisons. Moreover, AOPC scores are difficult to interpret in isolation without knowing the model-specific lower and upper limits. To address these issues, we propose a normalization approach, Normalized AOPC (NAOPC), enabling consistent cross-model evaluations and more meaningful interpretation of individual scores. Our experiments demonstrate that this normalization can radically change AOPC results, questioning the conclusions of earlier studies and offering a more robust framework for assessing feature attribution faithfulness. Our code is available at https://github.com/JoakimEdin/naopc.

When Backdoors Speak: Understanding LLM Backdoor Attacks Through Model-Generated Explanations

Huaizhi Ge, Yiming Li, Qifan Wang, Yongfeng Zhang, Ruixiang Tang

Large Language Models (LLMs) are known to be vulnerable to backdoor attacks, where triggers embedded in poisoned samples can maliciously alter LLMs' behaviors. In this paper, we move beyond attacking LLMs and instead examine backdoor attacks through the novel lens of natural language explanations. Specifically, we leverage LLMs' generative capabilities to produce human-readable explanations for their decisions, enabling direct comparisons between explanations for clean and poisoned samples. Our results show that backdoored models produce coherent explanations for clean inputs but diverse and logically flawed explanations for poisoned data, a pattern consistent across classification and generation tasks for different backdoor attacks. Further analysis reveals key insights into the explanation generation process. At the token level, explanation tokens associated with poisoned samples only appear in the final few transformer layers. At the sentence level, attention dynamics indicate that poisoned inputs shift attention away from the original input context during explanation generation. These findings enhance our understanding of backdoor mechanisms in LLMs and present a promising framework for detecting vulnerabilities through explainability.

Assessment and manipulation of latent constructs in pre-trained language models using psychometric scales

Maor Reuben, Ortal Slobodin, Idan-Chaim Cohen, Aviad Elyashar, Orna Braun-Lewensohn, Odeya Cohen, Rami Puzis

Human-like personality traits have recently been discovered in large language models, raising the hypothesis that their (known and as yet undiscovered) biases conform with human latent psychological constructs. While large conversational models may be tricked into answering psychometric questionnaires, the latent psychological constructs of thousands of simpler transformers, trained for other tasks, cannot be assessed because appropriate psychometric methods are currently lacking. Here, we show how standard psychological questionnaires can be reformulated into natural language inference prompts, and we provide a code library to support the psychometric assessment of arbitrary models. We demonstrate, using a sample of 88 publicly available models, the existence of human-like mental health-related constructs—including anxiety, depression, and the sense of coherence—which conform with standard theories in human psychology and show similar correlations and mitigation strategies. The ability to interpret and rectify the performance of language models by using psychological tools can boost the development of more explainable, controllable, and trustworthy models.

Cross-model Transferability among Large Language Models on the Platonic Representations of Concepts

Youcheng Huang, Chen Huang, Duanyu Feng, Wenqiang Lei, Jiancheng Lv

Understanding the inner workings of Large Language Models (LLMs) is a critical research frontier. Prior research has shown that a single LLM's concept representations can be captured as steering vectors (SVs), enabling the control of LLM behavior (e.g., towards generating harmful content). Our work takes a novel approach by exploring the intricate relationships between concept representations across different LLMs, drawing an intriguing parallel to Plato's Allegory of the Cave. In particular, we introduce a linear transformation method to bridge these representations and present three key findings: 1) Concept representations across different LLMs can be effectively aligned using simple linear transformations, enabling efficient cross-model transfer and behavioral control via SVs. 2) This linear transformation generalizes across concepts, facilitating alignment and control of SVs representing different concepts across LLMs. 3) A weak-to-strong transferability exists between LLM concept representations, whereby SVs extracted from smaller LLMs can effectively control the behavior of larger LLMs. Our code is provided in the supplementary file and will be openly released.

Interpret and Improve In-Context Learning via the Lens of Input-Label Mappings

Chenghao Sun, Zhen Huang, Yonggang Zhang, Le Lu, Houqiang Li, Xinmei Tian, Xu Shen, Jieping Ye

Large language models (LLMs) excel at downstream NLP tasks through in-context learning (ICL) with a few demonstrations of input–label pairs. However, the internal mechanisms behind ICL remain under-explored, particularly the mappings between inputs and labels. In this work, we reverse-engineer ICL by examining input-label mappings: what they are within LLMs, where they function, and how LLMs utilize them. (1) what: We discover input-label mappings stored within a few specific layers in the form of principal components (PCs), which capture human-interpretable and task-related words. (2) where: We propose a PC patching approach to identify the modules where input-label mappings function. Specifically, PC patching automatically crafts counterfactual representations using identified semantic PCs, rather than manually designing counterfactual text, to suppress the behavior related to LLM capability for ICL-related modules. Utilizing PC patching, we identify LLMs apply input-label mappings in a small fraction of attention heads. (3) how: We observe and verify that the identified key heads utilize input-label mappings from demonstrations to generate target labels for new queries. Based on these discoveries, we further show that precisely fine-tuning key ICL-related modules leads to significant improvements across diverse tasks.

Revisiting Epistemic Markers in Confidence Estimation: Can Markers Accurately Reflect Large Language Models' Uncertainty?

Jiayu Liu, Qing Zong, Weiqi Wang, Yangqiu Song

As large language models (LLMs) are increasingly used in high-stakes domains, accurately assessing their confidence is crucial. Humans typically express confidence through epistemic markers (e.g., "fairly confident") instead of numerical values. However, it remains unclear whether LLMs consistently use these markers to reflect their intrinsic confidence due to the difficulty of quantifying uncertainty associated with various markers. To address this gap, we first define ***marker confidence*** as the observed accuracy when a model employs an epistemic marker. We evaluate its stability across multiple question-answering datasets in both in-distribution and out-of-distribution settings for open-source and proprietary LLMs. Our results show that while markers generalize well within the same distribution, their confidence is inconsistent in out-of-distribution scenarios. These findings raise significant concerns about the reliability of epistemic markers for confidence estimation, underscoring the need for improved alignment between marker based confidence and actual model uncertainty. Our code is available at https://github.com/HKUST-KnowComp/MarCon.

ConSim: Measuring Concept-Based Explanations' Effectiveness with Automated Simulatability

Antonin Poché, Alon Jacovi, Agustin Martin Picard, Victor Boutin, Fanny Jourdan

Concept-based explanations work by mapping complex model computations to human-understandable concepts. Evaluating such explanations is very difficult, as it includes not only the quality of the induced space of possible concepts but also how effectively the chosen concepts are communicated to users. Existing evaluation metrics often focus solely on the former, neglecting the latter. We introduce an evaluation framework for measuring concept explanations via automated simulatability: a simulator's ability to predict the explained model's outputs based on the provided explanations. This approach accounts for both the concept space and its interpretation in an end-to-end evaluation. Human studies for simulatability are notoriously difficult to enact, particularly at the scale of a wide, comprehensive empirical evaluation (which is the subject of this work). We propose using large language models (LLMs) as simulators to approximate the evaluation and report various analyses to make such approximations reliable. Our method allows for scalable and consistent evaluation across various models and datasets. We report a comprehensive empirical evaluation using this framework and show that LLMs provide consistent rankings of explanation methods. Code available at Anonymous GitHub.

Enhancing Automated Interpretability with Output-Centric Feature Descriptions

Yoav Gur-Arieh, Roy Mayan, Chen Agassy, Atticus Geiger, Mor Geva

Automated interpretability pipelines generate natural language descriptions for the concepts represented by features in large language models (LLMs), such as "plants" or "the first word in a sentence". These descriptions are derived using inputs that activate the feature, which may be a dimension or a direction in the model's representation space. However, identifying activating inputs is costly, and the mechanistic role of a feature in model behavior is determined both by how inputs cause a feature to activate and by how feature activation affects outputs. Using steering evaluations, we reveal that current pipelines provide descriptions that fail to capture the causal effect of the feature on outputs. To fix this, we propose efficient, output-centric methods for automatically generating feature descriptions. These methods use the tokens weighted higher after feature stimulation or the highest weight tokens after applying the vocabulary "unembedding" head directly to the feature. Our output-centric descriptions better capture the causal effect of a feature on model outputs than input-centric descriptions, but combining the two leads to the best performance on both input and output evaluations. Lastly, we show that output-centric descriptions can be used to find inputs that activate features previously thought to be "dead".

Beyond Position: the emergence of wavelet-like properties in Transformers

Valeria Ruscio, Umberto Nanni, Fabrizio Silvestri

This paper studies how Transformer models with Rotary Position Embeddings (RoPE) develop emergent, wavelet-like properties that compensate for the positional encoding's theoretical limitations. Through an analysis spanning model scales, architectures, and training checkpoints, we show that attention heads evolve to implement multi-resolution processing analogous to wavelet transforms. We demonstrate that this scale-invariant behavior is unique to RoPE, emerges through distinct evolutionary phases during training, and statistically adheres to the fundamental uncertainty principle. Our findings suggest that the effectiveness of modern Transformers stems from their remarkable ability to spontaneously develop optimal, multi-resolution decompositions to address inherent architectural constraints.

Are the Hidden States Hiding Something? Testing the Limits of Factuality-Encoding Capabilities in LLMs

Giovanni Servedio, Alessandro De Bellis, Dario Di Palma, Vito Walter Anelli, Tommaso Di Noia

Factual hallucinations are a major challenge for Large Language Models (LLMs). They undermine reliability and user trust by generating inaccurate or fabricated content. Recent studies suggest that when generating false statements, the internal states of LLMs encode information about truthfulness. However, these studies often rely on synthetic datasets that lack realism, which limits generalization when evaluating the factual accuracy of text generated by the model itself. In this paper, we challenge the findings of previous work by investigating truthfulness encoding capabilities, leading to the generation of a more realistic and challenging dataset. Specifically, we extend previous work by introducing: (1) a strategy for sampling plausible true-false factoid sentences from tabular data and (2) a procedure for generating realistic, LLM-dependent true-false datasets from Question Answering collections. Our analysis of two open-source LLMs reveals that while the findings from previous studies are partially validated, generalization to LLM-generated datasets remains challenging. This study lays the groundwork for future research on factuality in LLMs and offers practical guidelines for more effective evaluation.

LLaMAs Have Feelings Too: Unveiling Sentiment and Emotion Representations in LLaMA Models Through Probing

Dario Di Palma, Alessandro De Bellis, Giovanni Servedio, Vito Walter Anelli, Fedelucio Narducci, Tommaso Di Noia

Large Language Models (LLMs) have rapidly become central to NLP, demonstrating their ability to adapt to various tasks through prompting techniques, including sentiment analysis. However, we still have a limited understanding of how these models capture sentiment-related information. This study probes the hidden layers of LLaMA models to pinpoint where sentiment features are most represented and to assess how this affects sentiment analysis. Using probe classifiers, we analyze sentiment encoding across layers and scales, identifying the layers and pooling methods that best capture sentiment signals. Our results show that sentiment information is most concentrated in mid-layers for binary polarity tasks, with detection accuracy increasing up to 14% over prompting techniques. Additionally, we find that in decoder-only models, the last token is not consistently the most informative for sentiment encoding. Finally, this approach enables sentiment tasks to be performed with memory requirements reduced by an average of 57%. These insights contribute to a broader understanding of sentiment in LLMs, suggesting layer-specific probing as an effective approach for sentiment tasks beyond prompting, with potential to enhance model utility and reduce memory requirements.

Drift: Enhancing LLM Faithfulness in Rationale Generation via Dual-Reward Probabilistic Inference

Jiazheng Li, Hanqi Yan, Yulan He

As Large Language Models (LLMs) are increasingly applied to complex reasoning tasks, achieving both accurate task performance and faithful explanations becomes crucial. However, LLMs often generate unfaithful explanations, partly because they do not consistently adhere closely to the provided context. Existing approaches to this problem either rely on superficial calibration methods, such as decomposed Chain-of-Thought prompting, or require costly retraining to improve model faithfulness. In this work, we propose a probabilistic inference paradigm that leverages task-specific and lookahead rewards to ensure that LLM-generated rationales are more faithful to model decisions and align better with input context. These rewards are derived from a domain-specific proposal distribution, allowing for optimized sequential Monte Carlo approximations. Our evaluations across three different reasoning tasks show that this method, which allows for controllable generation during inference, improves both accuracy and faithfulness of LLMs. This method offers a promising path towards making LLMs more reliable for reasoning tasks without sacrificing performance.

LLMs Struggle to Describe the Haystack without Human Help: A Social Science-Inspired Evaluation of Topic Models

Zongxia Li, Lorena Calvo-Bartolomé, Alexander Miserlis Hoyle, Paiheng Xu, Daniel Kofi Stephens, Juan Francisco Fung, Alden Dima, Jordan Lee Boyd-Graber

A common use of NLP is to facilitate the understanding of large document collections, with models based on Large Language Models (LLMs) replacing probabilistic topic models. Yet the effectiveness of LLM-based approaches in real-world applications remains under explored. This study measures the knowledge users acquire with topic models—including traditional, unsupervised and supervised LLM-based approaches—on two datasets. While LLM-based methods generate more human- readable topics and show higher average win probabilities than traditional models for data exploration, they produce overly generic topics for domain-specific datasets that do not easily allow users to learn much about the documents. Adding human supervision to LLM-based topic models improves data exploration by addressing hallucination and genericity but requires more human efforts. In contrast, traditional models like Latent Dirichlet Allocation (LDA) remain effective for exploration but are less user-friendly. This paper provides best practices—there is no one right model, the choice of models is situation-specific—and suggests potential improvements for scalable LLM- based topic models.

Improving Preference Extraction In LLMs By Identifying Latent Knowledge Through Classifying Probes

Sharan Maiya, Yinhong Liu, Ramit Debnath, Anna Korhonen

Large Language Models (LLMs) are often used as automated judges to evaluate text, but their effectiveness can be hindered by various unintentional biases. We propose using linear classifying probes, trained by leveraging differences between contrasting pairs of prompts, to directly access LLMs' latent knowledge and extract more accurate preferences. Through extensive experiments using models of varying size from four different families and six diverse datasets assessing text quality evaluation and common sense reasoning, we demonstrate that both supervised and unsupervised probing approaches consistently outperform traditional generation-based judgement while maintaining similar computational costs. These probes generalise under domain shifts and can even outperform finetuned evaluators with the same training data size. Our results suggest linear probing offers an accurate, robust and computationally efficient approach for LLM-as-judge tasks while providing interpretable insights into how models encode judgement-relevant knowledge. Our data and code will be openly released in the future.

DiffuseDef: Improved Robustness to Adversarial Attacks via Iterative Denoising

Zhenhao Li, Huichi Zhou, Marek Rei, Lucia Specia

Pretrained language models have significantly advanced performance across various natural language processing tasks. However, adversarial attacks continue to pose a critical challenge to system built using these models, as they can be exploited with carefully crafted adversarial texts. Inspired by the ability of diffusion models to predict and reduce noise in computer vision, we propose a novel and flexible adversarial defense method for language classification tasks, DiffuseDef, which incorporates a diffusion layer as a denoiser between the encoder and the classifier. The diffusion layer is trained on top of the existing classifier, ensuring seamless integration with any model in a plug-and-play manner. During inference, the adversarial hidden state is first combined with sampled noise, then denoised iteratively and finally ensembled to produce a robust text representation. By integrating adversarial training, denoising, and ensembling techniques, we show that DiffuseDef improves over existing adversarial defense methods and achieves state-of-the-art performance against common black-box and white-box adversarial attacks.

Curriculum Debiasing: Toward Robust Parameter-Efficient Fine-Tuning Against Dataset Biases

Mingyu Lee, Yeachan Kim, Wing-Lam Mok, SangKeun Lee

Parameter-efficient fine-tuning (PEFT) addresses the memory footprint issue of full fine-tuning by modifying only a subset of model parameters. However, on datasets exhibiting spurious correlations, we observed that PEFT slows down the model's convergence on unbiased examples, while the convergence on biased examples remains fast. This leads to the model's overfitting on biased examples, causing significant performance degradation in out-of-distribution (OOD) scenarios. Traditional debiasing methods mitigate this issue by emphasizing unbiased examples during training but often come at the cost of in-distribution (ID) performance drops. To address this trade-off issue, we propose a curriculum debiasing framework that presents examples in a \textit{biased-to-unbiased} order. Our framework initially limits the model’s exposure to unbiased examples, which are harder to learn, allowing it to first establish a foundation on easier-to-converge biased examples. As training progresses, we gradually increase the proportion of unbiased examples in the training set, guiding the model away from reliance on spurious correlations. Compared to the original PEFT methods, our method accelerates convergence on unbiased examples by approximately twofold and improves ID and OOD performance by 1.2% and 8.0%, respectively.

On Support Samples of Next Word Prediction

Yuqian Li, Yupei Du, Yufang Liu, Feifei Feng, Mou Xiao Feng, Yuanbin Wu

Language models excel in various tasks by making complex decisions, yet understanding the rationale behind these decisions remains a challenge. This paper investigates \emph{data-centric interpretability} in language models, focusing on the next-word prediction task. Using representer theorem, we identify two types of \emph{support samples}—those that either promote or deter specific predictions. Our findings reveal that being a support sample is an intrinsic property, predictable even before training begins. Additionally, while non-support samples are less influential in direct predictions, they play a critical role in preventing overfitting and shaping generalization and representation learning. Notably, the importance of non-support samples increases in deeper layers, suggesting their significant role in intermediate representation formation.These insights shed light on the interplay between data and model decisions, offering a new dimension to understanding language model behavior and interpretability.

Unveiling Attractor Cycles in Large Language Models: A Dynamical Systems View of Successive Paraphrasing

Zhilin Wang, Yafu Li, Jianhao Yan, Yu Cheng, Yue Zhang

Dynamical systems theory provides a framework for analyzing iterative processes and evolution over time. Within such systems, repetitive transformations can lead to stable configurations, known as attractors, including fixed points and limit cycles. Applying this perspective to large language models (LLMs), which iteratively map input text to output text, provides a principled approach to characterizing long-term behaviors. Successive paraphrasing serves as a compelling testbed for exploring such dynamics, as paraphrases re-express the same underlying meaning with linguistic variation. Although LLMs are expected to explore a diverse set of paraphrases in the text space, our study reveals that successive paraphrasing converges to stable periodic states, such as 2-period attractor cycles, limiting linguistic diversity. This phenomenon is attributed to the self-reinforcing nature of LLMs, as they iteratively favour and amplify certain textual forms over others. This pattern persists with increasing generation randomness or alternating prompts and LLMs. These findings underscore inherent constraints in LLM generative capability, while offering a novel dynamical systems perspective for studying their expressive potential.

CoT-ICL Lab: A Petri Dish for Studying Chain-of-Thought Learning from In-Context Demonstrations

Vignesh Kothapalli, Hamed Firooz, Maziar Sanjabi

We introduce CoT-ICL Lab, a framework and methodology to generate synthetic tokenized datasets and systematically study chain-of thought (CoT) in-context learning (ICL) in language models. CoT-ICL Lab allows fine grained control over the complexity of in-context examples by decoupling (1) the causal structure involved in chain token generation from (2) the underlying token processing functions. We train decoder-only transformers (up to 700M parameters) on these datasets and show that CoT accelerates the accuracy transition to higher values across model sizes. In particular, we find that model depth is crucial for leveraging CoT with limited in-context examples, while more examples help shallow models match deeper model performance. Additionally, limiting the diversity of token processing functions throughout training improves causal structure learning via ICL. We also interpret these transitions by analyzing transformer embeddings and attention maps. Overall, CoT-ICL Lab serves as a simple yet powerful testbed for theoretical and empirical insights into ICL and CoT in language models.

Towards Objective Fine-tuning: How LLMs' Prior Knowledge Causes Potential Poor Calibration?

Ziming Wang, Zeyu Shi, Haoyi Zhou, Shiqi Gao, Qingyun Sun, Jianxin Li

Fine-tuned Large Language Models (LLMs) often demonstrate poor calibration, with their confidence scores misaligned with actual performance. While calibration has been extensively studied in models trained from scratch, the impact of LLMs' prior knowledge on calibration during fine-tuning remains understudied. Our research reveals that LLMs' prior knowledge causes potential poor calibration due to the ubiquitous presence of known data in real-world fine-tuning, which appears harmful for calibration. Specifically, data aligned with LLMs' prior knowledge would induce overconfidence, while new knowledge improves calibration. Our findings expose a tension: LLMs' encyclopedic knowledge, while enabling task versatility, undermines calibration through unavoidable knowledge overlaps. To address this, we propose CogCalib, a cognition-aware framework that applies targeted learning strategies according to the model's prior knowledge. Experiments across 7 tasks using 3 LLM families prove that CogCalib significantly improves calibration while maintaining performance, achieving an average 57% reduction in ECE compared to standard fine-tuning in Llama3-8B. These improvements generalize well to out-of-domain tasks, enhancing the objectivity and reliability of domain-specific LLMs, and making them more trustworthy for critical human-AI interaction applications.

Circuit Compositions: Exploring Modular Structures in Transformer-Based Language Models

Philipp Mondorf, Sondre Wold, Barbara Plank

A fundamental question in interpretability research is to what extent neural networks, particularly language models, implement reusable functions through subnetworks that can be composed to perform more complex tasks. Recent advances in mechanistic interpretability have made progress in identifying circuits, the minimal computational subgraphs responsible for a model's behavior on specific tasks. However, most studies focus on identifying circuits for individual tasks without investigating how functionally similar circuits relate to each other. To address this gap, we study the modularity of neural networks by analyzing circuits for highly compositional subtasks within a transformer-based language model. Specifically, given a probabilistic context-free grammar, we identify and compare circuits responsible for ten modular string-edit operations. Our results indicate that functionally similar circuits exhibit both notable node overlap and cross-task faithfulness. Moreover, we demonstrate that the circuits identified can be reused and combined through set operations to represent more complex functional model capabilities.

Private Memorization Editing: Turning Memorization into a Defense to Strengthen Data Privacy in Large Language Models

Elena Sofia Ruzzetti, Giancarlo A. Xompero, Davide Venditti, Fabio Massimo Zanzotto

Large Language Models (LLMs) memorize, and thus, among huge amounts of uncontrolled data, may memorize Personally Identifiable Information (PII), which should not be stored and, consequently, not leaked. In this paper, we introduce Private Memorization Editing (PME), an approach for preventing private data leakage that turns an apparent limitation, that is, the LLMs' memorization ability, into a powerful privacy defense strategy. While attacks against LLMs have been performed exploiting previous knowledge regarding their training data, our approach aims to exploit the same kind of knowledge in order to make a model more robust. We detect a memorized PII and then mitigate the memorization of PII by editing a model knowledge of its training data. We verify that our procedure does not affect the underlying language model while making it more robust against privacy Training Data Extraction attacks. We demonstrate that PME can effectively reduce the number of leaked PII in a number of configurations, in some cases even reducing the accuracy of the privacy attacks to zero.

Does Time Have Its Place? Temporal Heads: Where Language Models Recall Time-specific Information

Yein Park, Chanwoong Yoon, Jungwoo Park, Minbyul Jeong, Jaewoo Kang

While the ability of language models to elicit facts has been widely investigated, how they handle temporally changing facts remains underexplored. We discover Temporal Heads, specific attention heads that primarily handle temporal knowledge, through circuit analysis. We confirm that these heads are present across multiple models, though their specific locations may vary, and their responses differ depending on the type of knowledge and its corresponding years. Disabling these heads degrades the model's ability to recall time-specific knowledge while maintaining its general capabilities without compromising time-invariant and question-answering performances. Moreover, the heads are activated not only numeric conditions (“In 2004”) but also textual aliases (“In the year ...”), indicating that they encode a temporal dimension beyond simple numerical representation. Furthermore, we expand the potential of our findings by demonstrating how temporal knowledge can be edited by adjusting the values of these heads.

Disentangling the Roles of Representation and Selection in Data Pruning

Yupei Du, Yingjin Song, Hugh Mee Wong, Daniil Ignatev, Albert Gatt, Dong Nguyen

Data pruning—selecting small but impactful subsets—offers a promising way to efficiently scale NLP model training. However, existing methods often involve many different design choices, which have not been systematically studied. This limits future developments. In this work, we decompose data pruning into two key components: data representation and selection algorithm, and systematically analyze their influence on selected instances. Our theoretical and empirical results highlight the crucial role of representations: better representations, e.g., training gradients, generally lead to better selected instances, regardless of the chosen selection algorithm. Furthermore, different selection algorithms excel in different settings, and none consistently outperform the others. Moreover, the selection algorithms do not always align with their intended objectives: for example, algorithms designed for the same objective can select drastically different instances, highlighting the need for careful evaluation.

Supervised Fine-Tuning Achieve Rapid Task Adaption Via Alternating Attention Head Activation Patterns

Yang Zhao, Li Du, Xiao Ding, Kai Xiong, Ting Liu, Bing Qin

LLMs' performance on complex tasks is still unsatisfactory. A key issue is that presently LLMs learn in a data-driven schema, while the instructions about these complex tasks are both scarce and hard to collect or construct. On the contrary, a prominent phenomenon is that LLMs can learn rather fast on simpler tasks with adequate prior knowledge captured during pretraining stage. Thus, if the prerequisite and mechanism of such rapid generalization could be elucidated, it could enhance the efficiency and effectiveness of the LLM's ability to learn complex tasks. Thus, in this paper, we employ a gradient-based method, to dissect the process that the SFT process adapts LLMs to downstream tasks via the perspective of attention patterns. We find that: (1) LLMs selectively activate task-specific attention heads during SFT; (2) activation patterns for complex tasks are combinations of basic task patterns; and (3) changes in a few parameters can significantly impact activation patterns after SFT on a small number of samples.Based on these insights, experiments are conducted to actually enhance the efficiency and effectiveness of SFT.

JoPA: Explaining Large Language Model's Generation via Joint Prompt Attribution

Yurui Chang, Bochuan Cao, Yujia Wang, Jinghui Chen, Lu Lin

Large Language Models (LLMs) have demonstrated impressive performances in complex text generation tasks. However, the contribution of the input prompt to the generated content still remains obscure to humans, underscoring the necessity of understanding the causality between input and output pairs. Existing works for providing prompt-specific explanation often confine model output to be classification or next-word prediction. Few initial attempts aiming to explain the entire language generation often treat input prompt texts independently, ignoring their combinatorial effects on the follow-up generation. In this study, we introduce a counterfactual explanation framework based on joint prompt attribution, JoPA, which aims to explain how a few prompt texts collaboratively influences the LLM's complete generation. Particularly, we formulate the task of prompt attribution for generation interpretation as a combinatorial optimization problem, and introduce a probabilistic algorithm to search for the casual input combination in the discrete space. We define and utilize multiple metrics to evaluate the produced explanations, demonstrating both the faithfulness and efficiency of our framework.

A Statistical and Multi-Perspective Revisiting of the Membership Inference Attack in Large Language Models

Bowen Chen, Namgi Han, Yusuke Miyao

The lack of data transparency in Large Language Models (LLMs) has highlighted the importance of Membership Inference Attack (MIA), which differentiates trained (member) and untrained (non-member) data. Though it shows success in previous studies, recent research reported a near-random performance in different settings, highlighting a significant performance inconsistency. We assume that a single setting doesn't represent the distribution of the vast corpora, causing members and non-members with different distributions to be sampled and causing inconsistency. In this study, instead of a single setting, we statistically revisit MIA methods from various settings with thousands of experiments for each MIA method, along with study in text feature, embedding, threshold decision, and decoding dynamics of members and non-members. We found that (1) MIA performance improves with model size and varies with domains, while most methods do not statistically outperform baselines, (2) Though MIA performance is generally low, a notable amount of differentiable member and non-member outliers exists and vary across MIA methods, (3) Deciding a threshold to separate members and non-members is an overlooked challenge, (4) Text dissimilarity and long text benefit MIA performance, (5) Differentiable or not is reflected in the LLM embedding, (6) Member and non-members show different decoding dynamics.

Mamba Knockout for Unraveling Factual Information Flow

Nir Endy, Idan Daniel Grosbard, Yuval Ran-Milo, Yonatan Slutzky, Itay Tshuva, Raja Giryes

This paper investigates the flow of factual information in Mamba State-Space Model (SSM)-based language models. We rely on theoretical and empirical connections to Transformer-based architectures and their attention mechanisms. Exploiting this relationship, we adapt attentional interpretability techniques originally developed for Transformers—specifically, the Attention Knockout methodology—to both Mamba-1 and Mamba-2. Using them we trace how information is transmitted and localized across tokens and layers, revealing patterns of subject-token information emergence and layer-wise dynamics. Notably, some phenomena vary between mamba models and Transformer based models, while others appear universally across all models inspected—hinting that these may be inherent to LLMs in general. By further leveraging Mamba's structured factorization, we disentangle how distinct "features" either enable token-to-token information exchange or enrich individual tokens, thus offering a unified lens to understand Mamba internal operations.

Rubrik's Cube: Testing a New Rubric for Evaluating Explanations on the CUBE dataset

Diana Galvan-Sosa, Gabrielle Gaudeau, Pride Kavumba, Yunmeng Li, Hongyi gu, Zheng Yuan, Keisuke Sakaguchi, Paula Buttery

The performance and usability of Large-Language Models (LLMs) are driving their use in explanation generation tasks. However, despite their widespread adoption, LLM explanations have been found to be unreliable, making it difficult for users to distinguish good from bad explanations. To address this issue, we present Rubrik's CUBE--an education-inspired rubric and a dataset of 26k explanations, written and later quality-annotated using the rubric by both humans and six open- and closed-source LLMs. The CUBE dataset focuses on two reasoning and two language tasks, providing the necessary diversity for us to effectively test our proposed rubric. Using Rubrik, we find that explanations are influenced by both task and perceived difficulty. Low quality stems primarily from a lack of conciseness in LLM-generated explanations, rather than cohesion and word choice. The full dataset, rubric, and code are available at https://github.com/RubriksCube/rubriks_cube.

Leveraging In-Context Learning for Political Bias Testing of LLMs

Patrick Haller, Jannis Vamvas, Rico Sennrich, Lena Ann Jäger

A growing body of work has been querying LLMs with political questions to evaluate their potential biases. However, this probing method has limited stability, making comparisons between models unreliable. In this paper, we argue that LLMs need more context. We propose a new probing task, Questionnaire Modeling (QM), that uses human survey data as in-context examples. We show that QM improves the stability of question-based bias evaluation, and demonstrate that it may be used to compare instruction-tuned models to their base versions. Experiments with LLMs of various sizes indicate that instruction tuning can indeed change the direction of bias. Furthermore, we observe a trend that larger models are able to leverage in-context examples more effectively, and generally exhibit smaller bias scores in QM. Data and code are publicly available.

Understanding the Dark Side of LLMs' Intrinsic Self-Correction

Qingjie Zhang, Di Wang, Haoting Qian, Yiming Li, Tianwei Zhang, Minlie Huang, Han Qiu

Intrinsic self-correction was initially proposed to improve LLMs' responses via feedback solely based on their inherent capability. However, recent works show that LLMs' intrinsic self-correction fails without oracle labels as feedback. In this paper, our research goal is to *interpret LLMs' intrinsic self-correction for different tasks, especially for those failure cases.* By including one simple task and three complex tasks with state-of-the-art (SOTA) LLMs like ChatGPT, Llama, and DeepSeek, we design three interpretation methods to reveal the dark side of LLMs' intrinsic self-correction. We identify intrinsic self-correction can (1) cause LLMs to waver both intermedia and final answers and lead to prompt bias on simple factual questions; (2) introduce human-like cognitive bias on complex tasks. In light of our findings, we also provide two simple yet effective strategies for alleviation: question repeating and supervised fine-tuning with a few samples. We open-source our work at https://x-isc.info/.

Inconsistent Tokenizations Cause Language Models to be Perplexed by Japanese Grammar

Andrew Gambardella, Takeshi Kojima, Yusuke Iwasawa, Yutaka Matsuo

Typical methods for evaluating the performance of language models evaluate their ability to answer questions accurately. These evaluation metrics are acceptable for determining the extent to which language models can understand and reason about text in a general sense, but fail to capture nuanced capabilities, such as the ability of language models to recognize and obey rare grammar points, particularly in languages other than English. We measure the perplexity of language models when confronted with the ``first person psych predicate restriction'' grammar point in Japanese. Weblab is the only tested open source model in the 7-10B parameter range which consistently assigns higher perplexity to ungrammatical psych predicate sentences than grammatical ones. We give evidence that Weblab's uniformly bad tokenization is a possible root cause for its good performance, and show that Llama 3's perplexity on grammatical psych predicate sentences can be reduced by orders of magnitude (28x difference) by restricting test sentences to those with uniformly well-behaved tokenizations. We show in further experiments on machine translation tasks that language models will use alternative grammar patterns in order to produce grammatical sentences when tokenization issues prevent the most natural sentence from being output.

Performance Gap in Entity Knowledge Extraction Across Modalities in Vision Language Models

Ido Cohen, Daniela Gottesman, Mor Geva, Raja Giryes

Vision-language models (VLMs) excel at extracting and reasoning about information from images. Yet, their capacity to leverage internal knowledge about specific entities remains underexplored. This work investigates the disparity in model performance when answering factual questions about an entity described in text versus depicted in an image. Our results reveal a significant accuracy drop — reaching 18\% for some models — when the entity is presented visually instead of textually. To study this gap we present PopVQA, a dataset which allows separating entity recognition and question answering, and use it to benchmark several models. We hypothesize that this decline arises from limitations in how information flows from image tokens to query tokens. Thus, we use mechanistic interpretability tools to reveal that, although image tokens are preprocessed by the vision encoder, meaningful information flow from these tokens occurs only in the much deeper layers. Furthermore, critical image processing happens in the language model's middle layers, allowing few layers for consecutive reasoning, highlighting a potential inefficiency in how the model utilizes its layers for reasoning. These insights shed light on the internal mechanics of VLMs and offer pathways for enhancing their reasoning capabilities. PopVQA can be found at https://huggingface.co/datasets/idoco/PopVQA.

Separating Tongue from Thought: Activation Patching Reveals Language-Agnostic Concept Representations in Transformers

Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, Robert West

A central question in multilingual language modeling is whether large language models (LLMs) develop a universal concept representation, disentangled from specific languages. In this paper, we address this question by analyzing latent representations (latents) during a word-translation task in transformer-based LLMs. We strategically extract latents from a source translation prompt and insert them into the forward pass on a target translation prompt. By doing so, we find that the output language is encoded in the latent at an earlier layer than the concept to be translated. Building on this insight, we conduct two key experiments. First, we demonstrate that we can change the concept without changing the language and vice versa through activation patching alone. Second, we show that patching with the mean representation of a concept across different languages does not affect the models' ability to translate it, but instead improves it. Finally, we generalize to multi-token generation and demonstrate that the model can generate natural language description of those mean representations. Our results provide evidence for the existence of language-agnostic concept representations within the investigated models.

Session 7: Interpretability and Analysis of Models for NLP (Tuesday 29 July, 10:30-12:00, Hall 5X)

Know-MRI: A Knowledge Mechanisms Revealer\&Interpreter for Large Language Models

Jiaxiang Liu, Boxuan Xing, Chenhao Yuan, ChenxiangZhang, Di Wu, Xiusheng Huang, Haida Yu, Chuhan Lang, Pengfei Cao, Jun Zhao, Kang Liu

As large language models (LLMs) continue to advance, there is a growing urgency to enhance the interpretability of their internal knowledge mechanisms. Consequently, many interpretation methods have emerged, aiming to unravel the knowledge mechanisms of LLMs from various perspectives. However, current interpretation methods differ in input data formats and interpreting outputs. The tools integrating these methods are only capable of supporting tasks with specific inputs, significantly constraining their practical applications. To address these challenges, we present an open-source **Know**ledge **M**echanisms **R**evealer\&**I**nterpreter (**Know-MRI**) designed to analyze the knowledge mechanisms within LLMs systematically. Specifically, we have developed an extensible core module that can automatically match different input data with interpretation methods and consolidate the interpreting outputs. It enables users to freely choose appropriate interpretation methods based on the inputs, making it easier to comprehensively diagnose the model's internal knowledge mechanisms from multiple perspectives. Our code is available at https://github.com/nlpkeg/Know-MRI. We also provide a demonstration video on https://youtu.be/OiP4T3MC518.

Revealing the Unwritten: Visual Investigation of Beam Search Trees to Address Language Model Prompting Challenges

Thilo Spinner, Rita Sevastjanova, Rebecca Kehlbeck, Tobias Stähle, Daniel A. Keim, Oliver Deussen, Andreas Spitz, Mennatallah El-Assady

The present popularity of generative language models has amplified interest in interactive methods to guide model outputs. Prompt refinement is considered one of the most effective means to influence output among these methods. We identify several challenges associated with prompting large language models, categorized into data- and model-specific, linguistic, and socio-linguistic challenges. A comprehensive examination of model outputs, including runner-up candidates and their corresponding probabilities, is needed to address these issues. The beam search tree, the prevalent algorithm to sample model outputs, can inherently supply this information. Consequently, we leverage an interactive visual method for investigating the beam search tree, facilitating analysis of the decisions made by the model during generation. Our explorative approach validates existing results and offers additional insights.

Dialz: A Python Toolkit for Steering Vectors

Zara Siddique, Liam Turner, Luis Espinosa-Anke

We introduce *Dialz*, a Python library for advancing research on steering vectors for open-source LMs. Steering vectors allow users to modify activations at inference time to amplify or weaken a 'concept', e.g. honesty or positivity, providing a more powerful alternative to prompting or fine-tuning. Dialz supports a diverse set of tasks, including creating contrastive pair datasets, computing and applying steering vectors, and visualizations. Unlike existing libraries, Dialz emphasizes modularity and usability, enabling both rapid prototyping and in-depth analysis. We demonstrate how Dialz can be used to reduce harmful outputs such as stereotypes, while also providing insights into model behaviour across different layers. We release Dialz with full documentation, tutorials, and support for popular open-source models to encourage further research in safe and controllable language generation. Dialz enables faster research cycles and facilitates insights into model interpretability, paving the way for safer, more transparent, and more reliable AI systems.

Session 7: Language Modeling (Tuesday 29 July, 10:30-12:00, Hall 4/5)

BadWindtunnel: Defending Backdoor in High-noise Simulated Training with Confidence Variance

Ruyi Zhang, Songlei Jian, Yusong Tan, Heng Gao, Haifang Zhou, Kai Lu

Current backdoor attack defenders in Natural Language Processing (NLP) typically involve data reduction or model pruning, risking losing crucial information. To address this challenge, we introduce a novel backdoor defender, i.e., BadWindtunnel, in which we build a high-noise simulated training environment, similar to the wind tunnel, which allows precise control over training conditions to model the backdoor learning behavior without affecting the final model. We also use the confidence variance as a learning behavior quantification metric in the simulated training, which is based on the characteristics of backdoor-poisoned data (shorted in poisoned data): higher learnability and robustness. In addition, we propose a two-step strategy to further model poisoned data, including target label identification and poisoned data revealing. Extensive experiments demonstrate BadWindtunnel's superiority, with a 21% higher average reduction in attack success rate than the second-best defender.

Tokenization is Sensitive to Language Variation

Anna Wegmann, Dong Nguyen, David Jurgens

Variation in language is ubiquitous and often systematically linked to regional, social, and contextual factors. Tokenizers split texts into smaller units and might behave differently for less common linguistic forms. This might affect downstream LLM performance differently on two types of tasks: Tasks where the model should be robust to language variation (e.g., for semantic tasks like NLI, labels do not depend on whether a text uses British or American spelling) and tasks where the model should be sensitive to language variation (e.g., for form-based tasks like authorship verification, labels depend on whether a text uses British or American spelling). We pre-train BERT base models for the popular Byte-Pair Encoding algorithm to investigate how key algorithmic design choices impact downstream models' performances: fitting corpus, pre-tokenizer and vocabulary size. We find that the best tokenizer varies on the two task types---with the pre-tokenizer having the biggest impact on performance. Further, we introduce a new approach to estimate tokenizer impact on downstream LLM performance, showing significant improvement over techniques like Rényi efficiency. We encourage more work on language variation and its relation to tokenizers and thus LLM performance.

DELMAN: Dynamic Defense Against Large Language Model Jailbreaking with Model Editing

Yi Wang, Fenghua Weng, Sibei Yang, Zhan Qin, Minlie Huang, Wenjie Wang

Large Language Models (LLMs) are widely applied in decision making, but their deployment is threatened by jailbreak attacks, where adversarial users manipulate model behavior to bypass safety measures. Existing defense mechanisms, such as safety fine-tuning and model editing, either require extensive parameter modifications or lack precision, leading to performance degradation on general tasks, which is unsuitable to post-deployment safety alignment. To address these challenges, we propose DELMAN (**D**ynamic **E**diting for **L**ong **M**s J**A**ilbreak Defe**N**se), a novel approach leveraging direct model editing for precise, dynamic protection against jailbreak attacks. DELMAN directly updates a minimal set of relevant parameters to neutralize harmful behaviors while preserving the model's utility. To avoid triggering a safe response in benign context, we incorporate KL-divergence regularization to ensure the updated model remains consistent with the original model when processing benign queries. Experimental results demonstrate that DELMAN outperforms baseline methods in mitigating jailbreak attacks while preserving the model’s utility, and adapts seamlessly to new attack instances, providing a practical and efficient solution for post-deployment model protection.

The Threat of PROMPTS in Large Language Models: A System and User Prompt Perspective

Zixuan Xia, Haifeng Sun, Jingyu Wang, Qi Qi, Huazheng Wang, Xiaoyuan Fu, Jianxin Liao

Prompts, especially high-quality ones, play an invaluable role in assisting large language models (LLMs) to accomplish various natural language processing tasks. However, carefully crafted prompts can also manipulate model behavior. Therefore, the security risks that "prompts themselves face" and those "arising from harmful prompts" cannot be overlooked and we define the Prompt Threat (PT) issues. In this paper, we review the latest attack methods related to prompt threats, focusing on prompt leakage attacks and prompt jailbreak attacks. Additionally, we summarize the experimental setups of these methods and explore the relationship between prompt threats and prompt injection attacks.

ACCESS DENIED INC: The First Benchmark Environment for Sensitivity Awareness

Dren Fazlija, Arkadij Orlov, Sandipan Sikdar

Large language models (LLMs) are increasingly becoming valuable to corporate data management due to their ability to process text from various document formats and facilitate user interactions through natural language queries. However, LLMs must consider the sensitivity of information when communicating with employees, especially given access restrictions. Simple filtering based on user clearance levels can pose both performance and privacy challenges. To address this, we propose the concept of sensitivity awareness (SA), which enables LLMs to adhere to predefined access rights rules. In addition, we developed a benchmarking environment called ACCESS DENIED INC to evaluate SA. Our experimental findings reveal significant variations in model behavior, particularly in managing unauthorized data requests while effectively addressing legitimate queries. This work establishes a foundation for benchmarking sensitivity-aware language models and provides insights to enhance privacy-centric AI systems in corporate environments.

Compute Optimal Scaling of Skills: Knowledge vs Reasoning

Nicholas Roberts, Niladri S. Chatterji, Sharan Narang, Mike Lewis, Dieuwke Hupkes

Scaling laws are a critical component of the LLM development pipeline, most famously as a way to forecast training decisions such as 'compute-optimally' trading-off parameter count and dataset size, alongside a more recent growing list of other crucial decisions. In this work, we ask whether compute-optimal scaling behaviour can be skill-dependent. In particular, we examine knowledge and reasoning-based skills such as knowledge-based QA and code generation, and we answer this question in the affirmative: scaling laws are skill-dependent. Next, to understand whether skill-dependent scaling is an artefact of the pretraining datamix, we conduct an extensive ablation of different datamixes and find that, also when correcting for datamix differences, knowledge and code exhibit fundamental differences in scaling behaviour. We conclude with an analysis of how our findings relate to standard compute-optimal scaling using a validation set, and find that a misspecified validation set can impact compute-optimal parameter count by nearly 50%, depending on its skill composition.

Robust Data Watermarking in Language Models by Injecting Fictitious Knowledge

Xinyue Cui, Johnny Wei, Swabha Swayamdipta, Robin Jia

Data watermarking in language models injects traceable signals, such as specific token sequences or stylistic patterns, into copyrighted text, allowing copyright holders to track and verify training data ownership. Previous data watermarking techniques primarily focus on effective memorization after pretraining, while overlooking challenges that arise in other stages of the LLM pipeline, such as the risk of watermark filtering during data preprocessing, or potential forgetting through post-training, or verification difficulties due to API-only access. We propose a novel data watermarking approach that injects coherent and plausible yet fictitious knowledge into training data using generated passages describing a fictitious entity and its associated attributes. Our watermarks are designed to be memorized by the LLM through seamlessly integrating in its training data, making them harder to detect lexically during preprocessing. We demonstrate that our watermarks can be effectively memorized by LLMs, and that increasing our watermarks' density, length, and diversity of attributes strengthens their memorization. We further show that our watermarks remain robust throughout LLM development, maintaining their effectiveness after continual pretraining and supervised finetuning. Finally, we show that our data watermarks can be evaluated even under API-only access via question answering.

Ranked Voting based Self-Consistency of Large Language Models

Weiqin Wang, Yile Wang, Hui Huang

Majority voting is considered an effective method to enhance chain-of-thought reasoning, as it selects the answer with the highest "self-consistency" among different reasoning paths (Wang et al., 2023). However, previous chain-of-thought reasoning methods typically generate only a single answer in each trial, thereby ignoring the possibility of other potential answers. As a result, these alternative answers are often overlooked in subsequent voting processes. In this work, we propose to generate ranked answers in each reasoning process and conduct ranked voting among multiple ranked answers from different responses, thereby making the overall self-consistency more reliable. Specifically, we use three ranked voting methods: Instant-runoff voting, Borda count voting, and mean reciprocal rank voting. We validate our methods on six datasets, including three multiple-choice and three open-ended question-answering tasks, using both advanced open-source and closed-source large language models. Extensive experimental results indicate that our proposed method outperforms the baselines, showcasing the potential of leveraging the information of ranked answers and using ranked voting to improve reasoning performance. Code and logs will be released.

Enhancing LLM Agent Safety via Causal Influence Prompting

Dongyoon Hahm, Woogyeol Jin, June Suk Choi, Sungsoo Ahn, Kimin Lee

As autonomous agents powered by large language models (LLMs) continue to demonstrate potential across various assistive tasks, ensuring their safe and reliable behavior is crucial for preventing unintended consequences. In this work, we introduce CIP, a novel technique that leverages causal influence diagrams (CIDs) to identify and mitigate risks arising from agent decision-making. CIDs provide a structured representation of cause-and-effect relationships, enabling agents to anticipate harmful outcomes and make safer decisions. Our approach consists of three key steps: (1) initializing a CID based on task specifications to outline the decision-making process, (2) guiding agent interactions with the environment using the CID, and (3) iteratively refining the CID based on observed behaviors and outcomes. Experimental results demonstrate that our method effectively enhances safety in both code execution and mobile device control tasks.

FANNO: Augmenting High-Quality Instruction Data with Open-Sourced LLMs Only

He Zhu, Yifan Ding, Yicheng Tao, Zhiwen Ruan, Yixia Li, Wenjia Zhang, Yun Chen, Guanhua Chen

Instruction tuning stands as a crucial advancement in leveraging large language models (LLMs) for enhanced task performance. However, the annotation of instruction datasets has traditionally been expensive and laborious, often relying on manual annotations or costly proprietary LLMs. Recent works explore approaches to synthesize data with open-sourced LLMs but require high-quality human-crafted seed data. In this work, we introduce \mname, an end-to-end framework to synthesize high-quality instruction data with open-sourced LLMs and sampled unlabeled documents, eliminating the necessity for seed data. Starting from diverse pre-screened documents, the framework synthesizes complex and diverse high-quality instruction and response pairs in different stages. We propose a tagging-based prompt method to generate diverse and complex seed data and a UCB-based approach to augment more instruction data with the seed data. A novel \textit{Think Different} prompt is proposed to address the distributional limitations of the seeds, further boosting the data diversity. Experiments prove that the \mname can generate diverse and complex high-quality data even with a opensource small teacher model. The synthesized instruction data demonstrates performance that is comparable to, or even surpasses, baseline annotation methods with proprietary LLMs or open-sourced LLMs while requiring fewer instruction data samples.

Process-based Self-Rewarding Language Models

Shimao Zhang, Xiao Liu, Xin Zhang, Junxiao Liu, Zheheng Luo, Shujian Huang, Yeyun Gong

Large Language Models have demonstrated outstanding performance across various downstream tasks and have been widely applied in multiple scenarios. Human-annotated preference data is used for training to further improve LLMs' performance, which is constrained by the upper limit of human performance. Therefore, Self-Rewarding method has been proposed, where LLMs generate training data by rewarding their own outputs. However, the existing self-rewarding paradigm is not effective in mathematical reasoning scenarios and may even lead to a decline in performance. In this work, we propose the Process-based Self-Rewarding pipeline for language models, which introduces long-thought reasoning, step-wise LLM-as-a-Judge, and step-wise preference optimization within the self-rewarding paradigm. Our new paradigm successfully enhances the performance of LLMs on multiple mathematical reasoning benchmarks through iterative Process-based Self-Rewarding, demonstrating the immense potential of self-rewarding to achieve artificial intelligence that may surpass human capabilities.

Privacy Ripple Effects from Adding or Removing Personal Information in Language Model Training

Jaydeep Borkar, Matthew Jagielski, Katherine Lee, Niloofar Mireshghallah, David A. Smith, Christopher A. Choquette-Choo

Due to the sensitive nature of personally identifiable information (PII), its owners may have the authority to control its inclusion or request its removal from large-language model (LLM) training. Beyond this, PII may be added or removed from training datasets due to evolving dataset curation techniques, because they were newly scraped for retraining, or because they were included in a new downstream fine-tuning stage. We find that the amount and ease of PII memorization is a dynamic property of a model that evolves throughout training pipelines and depends on commonly altered design choices. We characterize three such novel phenomena: (1) similar-appearing PII seen later in training can elicit memorization of earlier-seen sequences in what we call assisted memorization, and this is a significant factor (in our settings, up to 1/3); (2) adding PII can increase memorization of other PII; and (3) removing PII can lead to other PII being memorized.

ZeroDL: Zero-shot Distribution Learning for Text Clustering via Large Language Models

Hwiyeol Jo, Hyunwoo Lee, Taiwoo Park

The advancements in large language models (LLMs) have brought significant progress in NLP tasks. However, if a task cannot be fully described in prompts, the models could fail to carry out the task. In this paper, we propose a simple yet effective method to contextualize a task toward a LLM. The method utilizes (1) open-ended zero-shot inference from the entire dataset, (2) aggregating the inference results, and (3) finally incorporate the aggregated meta-information for the actual task. We show the effectiveness in text clustering tasks, empowering LLMs to perform text-to-text-based clustering and leading to improvements on several datasets. Lastly, we explore the generated class labels for clustering, showing how the LLM understands the task through data.

PromptWizard: Optimizing Prompts via Task-Aware, Feedback-Driven Self-Evolution

Eshaan Agarwal, Raghav Magazine, Joykirat Singh, Vivek Dani, Tanuja Ganu, Akshay Nambi

Large language models (LLMs) have transformed AI across diverse domains, with prompting being central to their success in guiding model outputs. However, manual prompt engineering is both labor-intensive and domain-specific, necessitating the need for automated solutions. We introduce PromptWizard, a novel, fully automated framework for discrete prompt optimization, utilizing a self-evolving, self-adapting mechanism. Through a feedback-driven critique and synthesis process, PromptWizard achieves an effective balance between exploration and exploitation, iteratively refining both prompt instructions and in-context examples to generate human-readable, task-specific prompts. This guided approach systematically improves prompt quality, resulting in superior performance across 45 tasks. PromptWizard excels even with limited training data, smaller LLMs, and various LLM architectures. Additionally, our cost analysis reveals a substantial reduction in API calls, token usage, and overall cost, demonstrating PromptWizard's efficiency, scalability, and advantages over existing prompt optimization strategies.

Extending LLM Context Window with Adaptive Grouped Positional Encoding: A Training-Free Method

Xinhao Xu, Jiaxin Li, Hui Chen, Zijia Lin, Jungong Han, Guiguang Ding

Processing long input remains a significant challenge for large language models (LLMs) due to the scarcity of large-scale long-context training data and the high computational cost of training models for extended context windows. In this paper, we propose **Ada**ptive **Gro**uped **P**ositional **E**ncoding (AdaGroPE), a training-free, plug-and-play method to enhance long-context understanding in existing LLMs. AdaGroPE progressively increases the reuse count of relative positions as the distance grows and dynamically adapts the positional encoding mapping to sequence length, thereby fully exploiting the range of pre-trained position embeddings. Its design is consistent with the principles of rotary position embedding (RoPE) and aligns with human perception of relative distance, enabling robust performance in real-world settings with variable-length inputs. Extensive experiments across various benchmarks demonstrate that our AdaGroPE consistently achieves state-of-the-art performance, surpassing baseline methods and even outperforming LLMs inherently designed for long-context processing on certain tasks.

ATLANTIS: Weak-to-Strong Learning via Importance Sampling

Yi Liu, Guoyin Wang, Shicheng Li, Feifan Song, Xu Sun

Supervised fine-tuning (SFT) enables large language models to align with training data for better performance in many aspects. Nevertheless, the gap between the distribution of current datasets from human annotations or model generations and the real-world data distribution heavily limits the capacities and potentials of models. As a result, we propose a new SFT technique, ATLANTIS, to bridge the gap. We adopt importance sampling to estimate the optimal data distribution in the real world from existing training datasets because the former is hard to sample from. Furthermore, we introduce an extra small model and reference model to estimate the sampling ratio through the probability gap between them. We evaluate our method with benchmarks in knowledge \\& understanding and preference aspects. The experiment results prove that ATLANTIS can bring consistent and significant improvements to models' performance. What's more, our method can be flexibly transferred among models with different structures. Our analyses demonstrate that our method is well-compatible with other SFT techniques to further enhance models' capacities and has great potential to be combined with existing training frameworks.

ObfusLM: Privacy-preserving Language Model Service against Embedding Inversion Attacks

Yu Lin, Ruining Yang, Yunlong Mao, Qizhi Zhang, Jue Hong, Quanwei Cai, Ye Wu, Huiqi Liu, zhiyu chen, Bing Duan, Sheng Zhong

As the rapid expansion of Machine Learning as a Service (MLaaS) for language models, concerns over the privacy of client inputs during inference or fine-tuning have correspondingly escalated. Recently, solutions have been proposed to safeguard client privacy by obfuscation techniques. However, the solutions incur notable decline in model utility and mainly focus on classification tasks, rendering them impractical for real-world applications. Moreover, recent studies reveal that these obfuscation, if not well designed, is susceptible to embedding inversion attacks (EIAs). In this paper, we devise ObfusLM, a privacy-preserving MLaaS framework for both classification and generation tasks. ObfusLM leverages a model obfuscation module to achieve privacy protection for both classification and generation tasks. Based on (k, \epsilon)-anonymity, ObfusLM includes novel obfuscation algorithms to reach provable security against EIAs. Extensive experiments show that ObfusLM outperforms existing works in utility by 10\% with a nearly 80\% resistance rate against EIAs.

APPL: A Prompt Programming Language for Harmonious Integration of Programs and Large Language Model Prompts

Honghua Dong, Qidong Su, Yubo Gao, Zhaoyu Li, Yangjun Ruan, Gennady Pekhimenko, Chris J. Maddison, Xujie Si

Large Language Models (LLMs) have become increasingly capable of handling diverse tasks with the aid of well-crafted prompts and integration of external tools, but as task complexity rises, the workflow involving LLMs can be complicated and thus challenging to implement and maintain. To address this challenge, we propose APPL, A Prompt Programming Language that acts as a bridge between computer programs and LLMs, allowing seamless embedding of prompts into Python functions, and vice versa. APPL provides an intuitive and Python-native syntax, an efficient parallelized runtime with asynchronous semantics, and a tracing module supporting effective failure diagnosis and replaying without extra costs. We demonstrate that APPL programs are intuitive, concise, and efficient through representative scenarios including Chain-of-Thought with self-consistency (CoT-SC) and ReAct tool-use agent. We further use LLMs to judge the language design between APPL and previous work, where the results indicate that codes written in APPL are more readable and intuitive. Our code, tutorial and documentation are available at https://github.com/appl-team/appl.

A Survey of Post-Training Scaling in Large Language Models

Hanyu Lai, Xiao Liu, Junjie Gao, Jiale Cheng, Zehan Qi, Yifan Xu, Shuntian Yao, Dan Zhang, Jinhua Du, Zhenyu Hou, Xin Lv, Minlie Huang, Yuxiao Dong, Jie Tang

Large language models (LLMs) have achieved remarkable proficiency in understanding and generating human natural languages, mainly owing to the "scaling law" that optimizes relationships among language modeling loss, model parameters, and pre-trained tokens. However, with the exhaustion of high-quality internet corpora and increasing computational demands, the sustainability of pre-training scaling needs to be addressed. This paper presents a comprehensive survey of post-training scaling, an emergent paradigm aiming to relieve the limitations of traditional pre-training by focusing on the alignment phase, which traditionally accounts for a minor fraction of the total training computation. Our survey categorizes post-training scaling into three key methodologies: Supervised Fine-tuning (SFT), Reinforcement Learning from Feedback (RLxF), and Test-time Compute (TTC). We provide an in-depth analysis of the motivation behind post-training scaling, the scalable variants of these methodologies, and a comparative discussion against traditional approaches. By examining the latest advancements, identifying promising application scenarios, and highlighting unresolved issues, we seek a coherent understanding and map future research trajectories in the landscape of post-training scaling for LLMs.

Pre-training Distillation for Large Language Models: A Design Space Exploration

Hao Peng, Xin Lv, Yushi Bai, Zijun Yao, Jiajie Zhang, Lei Hou, Juanzi Li

Knowledge distillation (KD) aims to transfer knowledge from a large teacher model to a smaller student model. Previous work applying KD in the field of large language models (LLMs) typically focused on the post-training phase, where the student LLM learns directly from instructions and corresponding responses generated by the teacher model. In this paper, we extend KD to the pre-training phase of LLMs, named pre-training distillation (PD). We first conduct a preliminary experiment using GLM-4-9B as the teacher LLM to distill a 1.9B parameter student LLM, validating the effectiveness of PD. Considering the key impact factors of distillation, we systematically explore the design space of pre-training distillation across four aspects: logits processing, loss selection, scaling law, and offline or online logits. We conduct extensive experiments to explore the design space of pre-training distillation and find better configurations and interesting conclusions, such as larger student LLMs generally benefiting more from pre-training distillation, while a larger teacher LLM does not necessarily guarantee better results. We hope our exploration of the design space will inform future practices in pre-training distillation.

LongReward: Improving Long-context Large Language Models with AI Feedback

Jiajie Zhang, Zhongni Hou, Xin Lv, Shulin Cao, Zhenyu Hou, Yilin Niu, Lei Hou, Yuxiao Dong, Ling Feng, Juanzi Li

Though significant advancements have been achieved in developing long-context large language models (LLMs), the compromised quality of LLM-synthesized data for supervised fine-tuning (SFT) often affects the long-context performance of SFT models and leads to inherent limitations. In principle, reinforcement learning (RL) with appropriate reward signals can further enhance models' capacities. However, how to obtain reliable rewards in long-context scenarios remains unexplored. To this end, we propose \textbf{LongReward}, a novel method that utilizes an off-the-shelf LLM to provide rewards for long-context model responses from four human-valued dimensions: helpfulness, logicality, faithfulness, and completeness, each with a carefully designed assessment pipeline. By combining LongReward and offline RL algorithm DPO, we are able to effectively improve long-context SFT models. Our experiments indicate that LongReward not only significantly improves models' long-context performance but also enhances their ability to follow short instructions. We also find that long-context DPO with LongReward and conventional short-context DPO can be used together without hurting either one's performance.

Positional Overload: Positional Debiasing and Context Window Extension for Large Language Models using Set Encoding

Lukas Kinder, Lukas Edman, Alexander Fraser, Tobias Käfer

Large Language Models (LLMs) typically track the order of tokens using positional encoding, which causes the following problems: positional bias, where the model is influenced by an ordering within the prompt, and a fixed context window, as models struggle to generalize to positions beyond those encountered during training. To address these limitations, we developed a novel method called \textit{set encoding}. This method allows multiple pieces of text to be encoded in the same position, thereby eliminating positional bias entirely. Another promising use case for set encoding is to increase the size of the input an LLM can handle. Our experiments demonstrate that set encoding allows an LLM to solve tasks with far more tokens than without set encoding. To our knowledge, set encoding is the first technique to effectively extend an LLM's context window without requiring any additional training.

Meta-Reflection: A Feedback-Free Reflection Learning Framework

Yaoke Wang, Yun Zhu, XintongBao, Wenqiao Zhang, Suyang Dai, kehan chen, Wenqiang Li, Gang Huang, Siliang Tang, Yueting Zhuang

Despite the remarkable capabilities of large language models (LLMs) in natural language understanding and reasoning, they often display undesirable behaviors, such as generating hallucinations and unfaithful reasoning. A prevalent strategy to mitigate these issues is the use of reflection, which refines responses through an iterative process. However, while promising, reflection heavily relies on high-quality external feedback and requires iterative multi-agent inference processes, thus hindering its practical application. In this paper, we propose Meta-Reflection, a novel feedback-free reflection mechanism that necessitates only a single inference pass without external feedback. Motivated by the human ability to remember and retrieve reflections from past experiences when encountering similar problems, Meta-Reflection integrates reflective insights into a codebook, allowing the historical insights to be stored, retrieved, and used to guide LLMs in problem-solving. To thoroughly investigate and evaluate the practicality of Meta-Reflection in real-world scenarios, we introduce an industrial e-commerce benchmark named E-commerce Customer Intent Detection. Extensive experiments conducted on both public datasets and the ECID benchmark highlight the effectiveness and efficiency of our proposed approach. Project is available at https://github.com/DCDmllm/Meta-Reflection

A Silver Bullet or a Compromise for Full Attention? A Comprehensive Study of Gist Token-based Context Compression

Chenlong Deng, Zhisong Zhang, Kelong Mao, Shuaiyi Li, Xinting Huang, Dong Yu, Zhicheng Dou

In this work, we provide an empirical investigation of gist-based context compression methods to improve context processing in large language models. We focus on two key questions: (1) How well can these methods replace full attention models? and (2) What potential failure patterns arise due to compression? Through extensive experiments, we show that while gist-based compression can achieve only slight performance loss on tasks like retrieval-augmented generation and long-document QA, it faces challenges in tasks like synthetic recall. Furthermore, we identify three key failure patterns: lost by the boundary, lost if surprise, and lost along the way. To mitigate these issues, we propose two effective strategies: fine-grained autoencoding, which enhances the reconstruction of original token information, and segment-wise token importance estimation, which adjusts optimization based on token dependencies. Our work provides valuable insights into the understanding of gist token-based context compression and offers practical strategies for improving compression capabilities.

Demons in the Detail: On Implementing Load Balancing Loss for Training Specialized Mixture-of-Expert Models

Zihan Qiu, Zeyu Huang, Bo Zheng, Kaiyue Wen, Zekun Wang, Rui Men, Ivan Titov, Dayiheng Liu, Jingren Zhou, Junyang Lin

This paper revisits the implementation of \textbf{L}oad-\textbf{B}alancing-\textbf{L}oss (LBL) when training Mixture-of-Experts (MoEs) models. Specifically, LBL for MoEs is defined as N_E \sum_{i=1}^{N_E} f_ip_i, where N_E is the total number of experts, f_i represents the frequency of expert i being selected, and p_i denotes the average gating score of the expert i. Existing MoE training frameworks usually employ the parallel training strategy so that f_i and the LBL are calculated within a \textbf{micro-batch} and averaged across parallel groups. However, a micro-batch for training billion-scale LLMs typically contains very few sequences, leading to the micro-batch LBL being almost at the sequence level, and the router is pushed to distribute the token evenly within each sequence. Under this strict constraint, even tokens from a domain-specific sequence (\textit{e.g.}, code) are uniformly routed to all experts, thereby inhibiting expert specialization. In this work, we propose calculating LBL using a \textbf{global-batch} to loose this constraint. Because a global-batch contains much more diverse sequences than a micro-batch, which will encourage load balance at the corpus level. Specifically, we introduce an extra communication step to synchronize f_i across micro-batches and then use it to calculate the LBL. Through experiments on training MoEs-based LLMs (up to \textbf{42.8B} parameters and \textbf{400B} tokens), we surprisingly find that the global-batch LBL strategy yields excellent performance gains in both pre-training perplexity and downstream tasks. Our analysis reveals that the global-batch LBL greatly improves the domain specialization of experts. \textit{Global-batch LBL is also used in Qwen3-MoEs.}

Hallucination Detox: Sensitivity Dropout (SenD) for Large Language Model Training

Shahrad Mohammadzadeh, Juan David Guerra, Marco Bonizzato, Reihaneh Rabbany, Golnoosh Farnadi

As large language models (LLMs) become increasingly prevalent, concerns about their reliability, particularly due to hallucinations - factually inaccurate or irrelevant outputs - have grown. Our research investigates the relationship between the uncertainty in training dynamics and the emergence of hallucinations. Using models from the Pythia suite and several hallucination detection metrics, we analyze hallucination trends and identify significant variance during training. To address this, we propose Sensitivity Dropout (SenD), a novel training protocol designed to reduce hallucination variance during training by deterministically dropping embedding indices with significant variability. In addition, we develop an unsupervised hallucination detection metric, Efficient EigenScore (EES), which approximates the traditional EigenScore in 2x speed. This metric is integrated into our training protocol, allowing SenD to be both computationally scalable and effective at reducing hallucination variance. SenD improves test-time reliability of Pythia and Meta's Llama models by up to 17% and enhances factual accuracy in Wikipedia, Medical, Legal, and Coding domains without affecting downstream task performance.

FocalPO: Enhancing Preference Optimizing by Focusing on Correct Preference Rankings

Tong Liu, Xiao Yu, Wenxuan Zhou, Jindong Gu, Volker Tresp

Efficient preference optimization algorithms such as Direct Preference Optimization (DPO) have become a popular approach in aligning large language models (LLMs) with human preferences. These algorithms implicitly treat the LLM as a reward model, and focus on training it to correct misranked preference pairs. However, recent work\citep{chen2024preference} empirically finds that DPO training \textit{rarely improves these misranked preference pairs}, despite its gradient emphasizing on these cases. We introduce FocalPO, a DPO variant that instead \textit{down-weighs} misranked preference pairs and prioritizes enhancing the model's understanding of pairs that it can already rank correctly. Inspired by Focal Loss used in vision tasks, FocalPO achieves this by adding a modulating factor to dynamically scale DPO loss. Our experiment demonstrates that FocalPO surpasses DPO and its variants on popular benchmarks like Alpaca Eval 2.0 and Arena-Hard using Mistral-Base-7B and Llama-3-Instruct-8B, with the introduced hyperparameter fixed. Additionally, we empirically reveals how FocalPO affects training on correct and incorrect sample groups, further underscoring its effectiveness.

SynthesizeMe! Inducing Persona-Guided Prompts for Personalized Reward Models in LLMs

Michael J Ryan, Omar Shaikh, Aditri Bhagirath, Daniel Frees, William Barr Held, Diyi Yang

Recent calls for pluralistic alignment of Large Language Models (LLMs) encourage adapting models to diverse user preferences. However, most prior work on personalized reward models heavily rely on additional identity information, such as demographic details or a predefined set of preference categories. To this end, we introduce SynthesizeMe, an approach to inducing synthetic user personas from user interactions for personalized reward modeling. SynthesizeMe first generates and verifies reasoning to explain user preferences, then induces synthetic user personas from that reasoning, and finally filters to informative prior user interactions in order to build personalized prompts for a particular user. We show that using SynthesizeMe induced prompts improves personalized LLM-as-a-judge accuracy by 4.4% on Chatbot Arena. Combining SynthesizeMe derived prompts with a reward model achieves top performance on PersonalRewardBench: a new curation of user-stratified interactions with chatbots collected from 854 users of Chatbot Arena and PRISM.

SoRFT: Issue Resolving with Subtask-oriented Reinforced Fine-Tuning

Zexiong Ma, Chao Peng, Pengfei Gao, Xiangxin Meng, Yanzhen Zou, Bing Xie

Mainstream issue-resolving frameworks predominantly rely on commercial models, leading to high costs and privacy concerns. Existing training approaches for issue resolving struggle with poor generalization and fail to fully leverage open-source development resources. We propose **S**ubtask-**o**riented **R**einforced **F**ine-**T**uning (**SoRFT**), a novel training approach to enhance the issue resolving capability of LLMs. We decomposes issue resolving into structured subtasks: file localization, function localization, line localization, and code edit generation. SoRFT consists of two training stages: (1) **rejection-sampled supervised fine-tuning**, Chain of Thought (CoT) data is filtered using ground-truth before fine-tuning the LLM, and (2) **rule-based reinforcement learning**, which leverages PPO with ground-truth based rewards. We evaluate the SoRFT-trained model on SWE-Bench Verified and SWE-Bench Lite, achieving state-of-the-art (SOTA) performance among open-source models (e.g., resolve 21.4% issues on SWE-Bench Verified with SoRFT-Qwen-7B). The experimental results demonstrate that SoRFT significantly enhances issue-resolving performance, improves model generalization, and provides a cost-efficient alternative to commercial models.

TreeRL: LLM Reinforcement Learning with On-Policy Tree Search

Zhenyu Hou, Ziniu Hu, Yujiang Li, Rui Lu, Jie Tang, Yuxiao Dong

Reinforcement learning (RL) with tree search has demonstrated superior performance in traditional reasoning tasks. Compared to conventional independent chain sampling strategies with outcome supervision, tree search enables better exploration of the reasoning space and provides dense, on-policy process rewards during RL training but remains under-explored in On-Policy LLM RL. We propose TreeRL, a reinforcement learning framework that directly incorporates on-policy tree search for RL training. Our approach includes intermediate supervision and eliminates the need for separate reward model training. Existing approaches typically train a separate process reward model, which can suffer from distribution mismatch and reward hacking. We also introduce a cost-effective tree search approach that achieves higher search efficiency under the same generation token budget by strategically branching from high-uncertainty intermediate steps rather than using random branching. Experiments on challenging math and code reasoning benchmarks demonstrate that TreeRL achieves superior performance compared to traditional ChainRL, highlighting the potential of tree search for LLM. TreeRL is open-sourced at .

LLaSE-G1: Incentivizing Generalization Capability for LLaMA-based Speech Enhancement

Boyi Kang, Xinfa Zhu, Zihan Zhang, Zhen Ye, Mingshuai Liu, Ziqian Wang, Yike Zhu, Guobin Ma, Jun Chen, Longshuai Xiao, CHAO WENG, Wei Xue, Lei Xie

Recent advancements in language models (LMs) have demonstrated strong capabilities in semantic understanding and contextual modeling, which have flourished in generative speech enhancement (SE). However, many LM-based SE approaches primarily focus on semantic information, often neglecting the critical role of acoustic information, which leads to acoustic inconsistency after enhancement and limited generalization across diverse SE tasks. In this paper, we introduce LLaSE-G1, a LLaMA-based language model that incentivizes generalization capabilities for speech enhancement. LLaSE-G1 offers the following key contributions: First, to mitigate acoustic inconsistency, LLaSE-G1 employs continuous representations from WavLM as input and predicts speech tokens from X-Codec2, maximizing acoustic preservation. Second, to promote generalization capability, LLaSE-G1 introduces dual-channel inputs and outputs, unifying multiple SE tasks without requiring task-specific IDs. Third, LLaSE-G1 outperforms prior task-specific discriminative and generative SE models, demonstrating scaling effects at test time and emerging capabilities for unseen SE tasks. Additionally, we release our code and models to support further research in this area.

DoMIX: An Efficient Framework for Exploiting Domain Knowledge in Fine-Tuning

Dohoon Kim, Donghun Kang, Taesup Moon

Domain-Adaptive Pre-training (DAP) has recently gained attention for its effectiveness in fine-tuning pre-trained models. Building on this, continual DAP has been explored to develop pre-trained models capable of incrementally incorporating different domain datasets. However, existing continual DAP methods face several limitations: (1) high computational cost and GPU memory usage during training; (2) sensitivity to incremental data order; and (3) providing a single, generalized model for all end tasks, which contradicts the essence of DAP. In this paper, we propose DoMIX, a novel approach that addresses these challenges by leveraging LoRA modules, a representative parameter-efficient fine-tuning (PEFT) method. Our approach enables efficient and parallel domain-adaptive pre-training that is robust to domain order and effectively utilizes accumulated knowledge to provide tailored pre-trained models for specific tasks. We also demonstrate that our method can be extended beyond the DAP setting to standard LLM fine-tuning scenarios. Code is available at https://github.com/dohoonkim-ai/DoMIX.

Learning Together to Perform Better: Teaching Small-Scale LLMs to Collaborate via Preferential Rationale Tuning

Sohan Patnaik, Milan Aggarwal, Sumit Bhatia, Balaji Krishnamurthy

LLMs such as GPT-4 have shown a remarkable ability to solve complex questions by generating step-by-step rationales. Prior works have utilized this capability to improve smaller and cheaper LMs (say, with 7B parameters). However, various practical constraints, such as copyright and legal issues, owing to lack of transparency in the pre-training data of large (often closed) models, prevent their use in commercial settings. Little focus has been given to improving the innate reasoning ability of smaller models without distilling information from larger LLMs. To address this, we propose COLLATE, a trainable framework that tunes a (small) LLM to generate those outputs from a pool of diverse rationales that selectively improves the downstream task. COLLATE enforces multiple instances of the same LLM to exhibit distinct behavior and employs them to generate rationales to obtain diverse outputs. The LLM is then tuned via preference optimization to choose the candidate rationale which maximizes the likelihood of ground-truth answer. COLLATE outperforms several trainable and prompting baselines on 5 datasets across 3 domains - maths problem solving, natural language inference, and commonsense reasoning. We show the efficacy of COLLATE on LLMs from different model families across varying parameter scales (1B to 8B) and demonstrate the benefit of multiple rationale providers guided by the end task through ablations. Code is released here (https://github.com/Sohanpatnaik106/collate).

Agentic Reward Modeling: Integrating Human Preferences with Verifiable Correctness Signals for Reliable Reward Systems

Hao Peng, Yunjia Qi, Xiaozhi Wang, Zijun Yao, Bin Xu, Lei Hou, Juanzi Li

Reward models (RMs) are crucial for the training and inference-time scaling up of large language models (LLMs). However, existing reward models primarily focus on human preferences, neglecting verifiable correctness signals which have shown strong potential in training LLMs. In this paper, we propose agentic reward modeling, a reward system that combines reward models with verifiable correctness signals from different aspects to provide reliable rewards. We empirically implement a reward agent, named RewardAgent, that combines human preference rewards with two verifiable signals: factuality and instruction following, to provide more reliable rewards. We conduct comprehensive experiments on existing reward model benchmarks and inference-time best-of-n searches on real-world downstream tasks. RewardAgent significantly outperforms vanilla reward models, demonstrating its effectiveness. We further construct training preference pairs using RewardAgent and train an LLM with the DPO objective, achieving superior performance on various NLP benchmarks compared to conventional reward models. Our codes are publicly released to facilitate further research.

DualGuard: A Parameter Space Transformation Approach for Bidirectional Defense in Split-Based LLM Fine-Tuning

Zihan Liu, Yizhen Wang, Rui Wang, Sai Wu

Integrating split learning with large language model fine-tuning (LLM-FT) enables secure collaboration between a trusted local client and a well-equipped remote server, but it is vulnerable to data reconstruction attacks (DRAs) that exploit transmitted activations and gradients. Current defense methods, like adding noise to activations or gradients, often sacrifice task-specific model performance under strict privacy constraints. This paper introduces DualGuard, a bidirectional defense mechanism against DRAs for split-based LLM-FT. DualGuard proposes a local warm-up parameter space transformation to alter client-side model parameters before training, using multi-task learning to strike a balance between privacy protection and model performance. Additionally, a global fine-tuning parameter space retention strategy prevents the model from reverting to vulnerable states during formal fine-tuning. Experiments show that DualGuard outperforms current defense methods against various DRAs, while maintaining task performance. Our code will be made publicly available.

AgentRM: Enhancing Agent Generalization with Reward Modeling

Yu Xia, Jingru Fan, Weize Chen, Siyu Yan, Xin Cong, Zhong Zhang, Yaxi Lu, Yankai Lin, Zhiyuan Liu, Maosong Sun

Existing LLM-based agents have achieved strong performance on held-in tasks, but their generalizability to unseen tasks remains poor. Hence, some recent work focus on fine-tuning the policy model with more diverse tasks to improve the generalizability. In this work, we find that finetuning a reward model to guide the policy model is more robust than directly finetuning the policy model. Based on this finding, we propose AgentRM, a 8B generalizable reward model, to guide the policy model for effective test-time search. We comprehensively investigate three approaches to construct the reward model, including explicit reward modeling, implicit reward modeling and LLM-as-a-judge. We then use AgentRM to guide the answer generation with Best-of-N sampling and beam search. We show that AgentRM is robust to paraphrasings of task instructions and can generalize to unseen tasks that require novel optimal behavior. Through extensive evaluation across nine tasks spanning four categories, AgentRM enhances the non-finetuned 8B policy model by 8.8 points on average, surpassing the top general agent by 4.0. Moreover, it demonstrates weak-to-strong generalization, yielding greater improvement on more powerful policy models. As for the specializability, AgentRM can also boost a finetuned policy model and outperform the top specialized agent by 11.4 on three held-in tasks. Further analysis verifies its effectiveness in test-time scaling. We release the code and data at https://github.com/thunlp/AgentRM.

Understanding Silent Data Corruption in LLM Training

Jeffrey Jian Ma, Hengzhi Pei, Leonard Lausen, George Karypis

As the scale of training large language models (LLMs) increases, one emergent failure is silent data corruption (SDC), where hardware produces incorrect computations without explicit failure signals. In this work, we are the first to investigate the impact of real-world SDCs on LLM training by comparing model training between healthy production nodes and unhealthy nodes exhibiting SDCs. With the help from a cloud computing platform, we access the unhealthy nodes that were swept out from production by automated fleet management. Using deterministic execution via XLA compiler and our proposed synchronization mechanisms, we isolate and analyze the impact of SDC errors on these nodes at three levels: at each submodule computation, at a single optimizer step, and at a training period. Our results reveal that the impact of SDCs on computation varies on different unhealthy nodes. Although in most cases the perturbations from SDCs on submodule computation and gradients are relatively small, SDCs can lead models to converge to different optima with different weights and even cause spikes in the training loss. Our analysis sheds light on further understanding and mitigating the impact of SDCs.

LESA: Learnable LLM Layer Scaling-Up

Yifei Yang, zouying cao, Xinbei Ma, Yao Yao, Zhi Chen, Libo Qin, hai zhao

Training Large Language Models (LLMs) from scratch requires immense computational resources, making it prohibitively expensive. Model scaling-up offers a promising solution by leveraging the parameters of smaller models to create larger ones. However, existing depth scaling-up methods rely on empirical heuristic rules for layer duplication, which result in poorer initialization and slower convergence during continual pre-training. We propose \textbf{LESA}, a novel learnable method for depth scaling-up. By concatenating parameters from each layer and applying Singular Value Decomposition, we uncover latent patterns between layers, suggesting that inter-layer parameters can be learned. LESA uses a neural network to predict the parameters inserted between adjacent layers, enabling better initialization and faster training. Experiments show that LESA outperforms existing baselines, achieving superior performance with less than half the computational cost during continual pre-training. Extensive analyses demonstrate its effectiveness across different model sizes and tasks.

Synthesizing Post-Training Data for LLMs through Multi-Agent Simulation

Shuo Tang, Xianghe Pang, Zexi Liu, Bohan Tang, Rui Ye, Tian Jin, Xiaowen Dong, Yanfeng Wang, Siheng Chen

Post-training is essential for enabling large language models (LLMs) to follow human instructions. However, its effectiveness depends on high-quality instruction data, which is challenging to obtain in the real world due to privacy concerns, data scarcity, and high annotation costs. To fill this gap, inspired by the recent success of using LLMs to simulate human society, we propose MATRIX, a multi-agent simulator that automatically generates diverse text-based scenarios, capturing a wide range of real-world human needs in a realistic and scalable manner. Leveraging these outputs, we introduce a novel scenario-driven instruction generator MATRIX-Gen for controllable and highly realistic data synthesis. Extensive experiments demonstrate that our framework effectively generates both general and domain-specific data. On AlpacaEval 2 and Arena-Hard benchmarks, Llama-3-8B-Base, post-trained on datasets synthesized by MATRIX-Gen with just 20K instruction-response pairs, outperforms Meta's Llama-3-8B-Instruct model, which was trained on over 10M pairs.

Small Changes, Big Impact: How Manipulating a Few Neurons Can Drastically Alter LLM Aggression

Jaewook Lee, Junseo Jang, Oh-Woog KWON, Harksoo Kim

Recent remarkable advances in Large Language Models (LLMs) have led to innovations in various domains such as education, healthcare, and finance, while also raising serious concerns that they can be easily misused for malicious purposes. Most previous research has focused primarily on observing how jailbreak attack techniques bypass safety mechanisms like Reinforcement Learning through Human Feedback (RLHF). However, whether there are neurons within LLMs that directly govern aggression has not been sufficiently investigated. To fill this gap, this study identifies specific neurons ("aggression neurons") closely related to the expression of aggression and systematically analyzes how manipulating them affects the model's overall aggression. Specifically, using a large-scale synthetic text corpus (aggressive and non-aggressive), we measure the activation frequency of each neuron, then apply masking and activation techniques to quantitatively evaluate changes in aggression by layer and by manipulation ratio. Experimental results show that, in all models, manipulating only a small number of neurons can increase aggression by up to 33%, and the effect is even more extreme when aggression neurons are concentrated in certain layers. Moreover, even models of the same scale exhibit nonlinear changes in aggression patterns, suggesting that simple external safety measures alone may not be sufficient for complete defense.

SudoLM: Learning Access Control of Parametric Knowledge with Authorization Alignment

Qin Liu, Fei Wang, Chaowei Xiao, Muhao Chen

Existing preference alignment is a one-size-fits-all alignment mechanism, where the part of the large language model (LLM) parametric knowledge with non-preferred features is uniformly blocked to all the users. However, this part of knowledge can be useful to advanced users whose expertise qualifies them to handle these information. The one-size-fits-all alignment mechanism undermines LLM's utility for these qualified users. To address this problem, we propose SudoLM, a framework that lets LLMs learn access control over specific parametric knowledge for users with different credentials via authorization alignment. SudoLM allows authorized users to unlock their access to all the parametric knowledge with an assigned Sudo key while blocking access to non-qualified users. Experiments on two application scenarios demonstrate that SudoLM effectively controls the user's access to the parametric knowledge and maintains its general utility.

Learning to Reason Over Time: Timeline Self-Reflection for Improved Temporal Reasoning in Language Models

Adrián Bazaga, Rexhina Blloshmi, Bill Byrne, Adrià de Gispert

Large Language Models (LLMs) have emerged as powerful tools for generating coherent text, understanding context, and performing reasoning tasks. However, they struggle with temporal reasoning, which requires processing time-related information such as event sequencing, durations, and inter-temporal relationships. These capabilities are critical for applications including question answering, scheduling, and historical analysis. In this paper, we introduce TISER, a novel framework that enhances the temporal reasoning abilities of LLMs through a multi-stage process that combines timeline construction with iterative self-reflection. Our approach leverages test-time scaling to extend the length of reasoning traces, enabling models to capture complex temporal dependencies more effectively. This strategy not only boosts reasoning accuracy but also improves the traceability of the inference process. Experimental results demonstrate state-of-the-art performance across multiple benchmarks, including out-of-distribution test sets, and reveal that TISER enables smaller open-source models to surpass larger closed-weight models on challenging temporal reasoning tasks.

Learn to Memorize: Scalable Continual Learning in Semiparametric Language Models with Mixture-of-Neighbors Induction Memory

Guangyue Peng, Tao Ge, Wen Luo, Wei Li, Houfeng Wang

Semiparametric language models (LMs) have shown promise in various Natural Language Processing (NLP) tasks. However, they utilize non-parametric memory as static storage, which lacks learning capability and remains disconnected from the internal information flow of the parametric models, limiting scalability and efficiency. Based on recent interpretability theories of LMs, we reconceptualize the non-parametric memory represented by kNN-LM as a learnable Mixture-of-Neighbors Induction Memory (MoNIM), which synergizes the induction capabilities of attention heads with the memorization strength of feed-forward networks (FFN). By integrating into the model's information flow, MoNIM functions as an FFN-like bypass layer within the Transformer architecture, enabling effective learning of new knowledge. Extensive experiments demonstrate that MoNIM is a retentive and scalable continual learner in both data- and model-wise, enhancing the scalability and continual learning performance of semiparametric LMs.\footnote{Code is publicly available at .}

MEraser: An Effective Fingerprint Erasure Approach for Large Language Models

Jingxuan Zhang, Zhenhua Xu, Rui Hu, Wenpeng Xing, Xuhong Zhang, Meng Han

Large Language Models (LLMs) have become increasingly prevalent across various sectors, raising critical concerns about model ownership and intellectual property protection. Although backdoor-based fingerprinting has emerged as a promising solution for model authentication, effective attacks for removing these fingerprints remain largely unexplored. Therefore, We present Mismatched Eraser (MEraser), a novel method for effectively removing backdoor-based fingerprints from LLMs while maintaining model performance. Our approach leverages a two-phase fine-tuning strategy utilizing carefully constructed mismatched and clean datasets. Through extensive evaluation across multiple LLM architectures and fingerprinting methods, we demonstrate that MEraser achieves complete fingerprinting removal while maintaining model performance with minimal training data of fewer than 1,000 samples. Furthermore, we introduce a transferable erasure mechanism that enables effective fingerprinting removal across different models without repeated training. In conclusion, our approach provides a practical solution for fingerprinting removal in LLMs, reveals critical vulnerabilities in current fingerprinting techniques, and establishes comprehensive evaluation benchmarks for developing more resilient model protection methods in the future.

LSSF: Safety Alignment for Large Language Models through Low-Rank Safety Subspace Fusion

Guanghao Zhou, Panjia Qiu, Cen Chen, Hongyu Li, Jason Chu, Xin Zhang, JUN ZHOU

The safety mechanisms of large language models (LLMs) exhibit notable fragility, as even fine-tuning on datasets without harmful content may still undermine their safety capabilities. Meanwhile, existing safety alignment methods predominantly rely on the fine-tuning process, which inadvertently leads to the increased complexity and computational resources required. To address these issues, we introduce LSSF, a novel safety re-alignment framework with \underline{L}ow-Rank \underline{S}afety \underline{S}ubspace \underline{F}usison. Our proposed method exploits the low-rank characteristics of safety information in LLMs by constructing a low-rank projection matrix to extract the principal components of safety vectors. Notably, this projection matrix represents the low-rank safety subspace of the LLMs, which we have observed to remain stable during fine-tuning process and is isolated from the model's general capabilities. These principal components are used to effectively restore safety alignment when combined with fine-tuned LLMs through linear arithmetic. Additionally, to account for the varying encoding densities of safety information across different layers of LLMs, we propose a novel metric called safety singular value entropy. This metric quantifies the encoding density and allows for the dynamic computation of the safety-critical rank for each safety vector. Extensive experiments demonstrate that our proposed post-hoc alignment method can effectively restore the safety alignment of fine-tuned models with minimal impact on their performance on downstream tasks.

Beyond Text Compression: Evaluating Tokenizers Across Scales

Jonas F. Lotz, António V. Lopes, Stephan Peitz, Hendra Setiawan, Leonardo Emili

The choice of tokenizer can profoundly impact language model performance, yet accessible and reliable evaluations of tokenizer quality remain an open challenge. Inspired by scaling consistency, we show that smaller models can accurately predict significant differences in tokenizer impact on larger models at a fraction of the compute cost. By systematically evaluating both English-centric and multilingual tokenizers, we find that tokenizer choice has negligible effects on tasks in English but results in consistent performance differences in multilingual settings. We propose new intrinsic tokenizer metrics inspired by Zipf's law that correlate more strongly with downstream performance than text compression when modeling unseen languages. By combining several metrics to capture multiple aspects of tokenizer behavior, we develop a reliable framework for intrinsic tokenizer evaluations. Our work offers a more efficient path to informed tokenizer selection in future language model development.

Emergent Abilities of Large Language Models under Continued Pre-training for Language Adaptation

Ahmed Elhady, Eneko Agirre, Mikel Artetxe

Continued pretraining (CPT) is a popular approach to adapt existing large language models (LLMs) to new languages. When doing so, it is common practice to include a portion of English data in the mixture, but its role has not been carefully studied to date. In this work, we show that including English does not impact validation perplexity, yet it is critical for the emergence of downstream capabilities in the target language. We introduce a language-agnostic benchmark for in-context learning (ICL), which reveals catastrophic forgetting early on CPT when English is not included. This in turn damages the ability of the model to generalize to downstream prompts as measured by perplexity, even if it does not manifest in terms of accuracy until later in training, and can be tied to a big shift in the model parameters. Based on these insights, we introduce curriculum learning and exponential moving average (EMA) of weights as effective alternatives to mitigate the need for English. All in all, our work sheds light into the dynamics by which emergent abilities arise when doing CPT for language adaptation, and can serve as a foundation to design more effective methods in the future.

Session 7: Language Modeling (Tuesday 29 July, 10:30-12:00, Hall 5X)

GigaChat Family: Efficient Russian Language Modeling Through Mixture of Experts Architecture

Mamedov Valentin Yurievich, Evgenii Kosarev, Gregory Leleytner, Ilya Shchuckin, Valeriy Berezovskiy, Daniil Smirnov, Dmitry Kozlov, Sergei Averkiev, Lukyanenko Ivan, Aleksandr Proshunin, Ainur Israfilova, Ivan Baskov, Artem Chervyakov, Emil Shakirov, Mikhail Kolesov, Daria Khomich, , Sergei Porkhun, Yury Fedorov, Oleg Kutuzov, Polina Kudriavtseva, Sofiia Soldatova, Kolodin Egor, Stanislav Pyatkin, DZMITRY MENSHYKH, Grafov Sergei IUrevich, Eldar Damirov, Karlov Vladimir, Ruslan Gaitukiev, Arkadiy Shatenov, Alena Fenogenova, Nikita Savushkin, Fedor Minkin

Generative large language models (LLMs) have become crucial for modern NLP research and applications across various languages. However, the development of foundational models specifically tailored to the Russian language has been limited, primarily due to the significant computational resources required. This paper introduces the GigaChat family of Russian LLMs, available in various sizes, including base models and instruction-tuned versions. We provide a detailed report on the model architecture, pre-training process, and experiments to guide design choices. In addition, we evaluate their performance on Russian and English benchmarks and compare GigaChat with multilingual analogs. The paper presents a system demonstration of the top-performing models accessible via an API, a Telegram bot, and a Web interface. Furthermore, we have released three open GigaChat models in open source, with the aim of expanding NLP research opportunities and supporting the development of industrial solutions for the Russian language.

Session 7: Machine Learning for NLP (Tuesday 29 July, 10:30-12:00, Hall 4/5)

Exploring Graph Representations of Logical Forms for Language Modeling

Michael Sullivan

We make the case for language models over logical forms (LFLMs), arguing that such models are more data-efficient than their textual counterparts. To that end, we introduce the Graph-based Formal-Logical Distributional Semantics (GFoLDS) prototype, a pretrained LM over graph representations of logical forms, as a proof-of-concept of LFLMs. Using GFoLDS, we present strong experimental evidence that LFLMs can leverage the built-in, basic linguistic knowledge inherent in such models to immediately begin learning more complex patterns. On downstream tasks, we show that GFoLDS vastly outperforms textual, transformer LMs (BERT) pretrained on the same data, indicating that LFLMs can learn with substantially less data than models over plain text. Furthermore, we show that the performance of this model is likely to scale with additional parameters and pretraining data, suggesting the viability of LFLMs in real-world applications.

SGDPO: Self-Guided Direct Preference Optimization for Language Model Alignment

Wenqiao Zhu, Ji Liu, Lulu Wang, Jun Wu, Yulun Zhang

Direct Preference Optimization (DPO) is broadly utilized for aligning Large Language Models (LLMs) with human values because of its flexibility. Despite its effectiveness, it has been observed that the capability of DPO to generate human-preferred response is limited and the results of DPO are far from resilient. To address these limitations, in this paper we propose a novel Self-Guided Direct Preference Optimization algorithm, i.e., SGDPO, which incorporates a pilot term to steer the gradient flow during the optimization process, allowing for fine-grained control over the updates of chosen and rejected rewards. We provide a detailed theoretical analysis of our proposed method and elucidate its operational mechanism. Furthermore, we conduct comprehensive experiments on various models and benchmarks. The extensive experimental results demonstrate the consistency between the empirical results and our theoretical analysis and confirm the effectiveness of our proposed approach (up to 9.19% higher score).

Maximum Score Routing For Mixture-of-Experts

Bowen Dong, Yilong Fan, Yutao Sun, Zhenyu Li, Tengyu Pan, zhou Xun, Jianyong Wang

Routing networks in sparse activated mixture-of-experts dynamically allocate input tokens to top-k experts through differentiable sparse transformations, which can expand neural networks while maintaining computational efficiency. Traditional MoE networks use an expert capacity constraint to achieve GPU-friendly computing. However, when the token is routed to experts with saturated capacity, token dropping occurs, and the padding operation of underutilized experts leads to low hardware efficiency. Removing the capacity constraint can also result in low computational efficiency and decreased load balancing. We propose **Maximum Score Routing (MaxScore)** in this work, a new MoE routing paradigm developed by modeling MoE routing as the minimum cost maximum flow problem and integrating a Soft-Topk operator. MaxScore retains the expert capacity constraint, and under the same FLOPs, compared to the constrained and unconstrained baselines, MaxScore exhibits lower training losses and higher evaluation scores.

Prediction-Augmented Generation for Automatic Diagnosis Tasks

Chan-Yang Ju, Dong-Ho Lee

Most Large language models (LLMs) adopt an autoregressive architecture, predicting the next word token based on the preceding context. While this approach is robust for language generation tasks such as writing and summarization, it has limitations for high-level reasoning tasks, such as prediction and decision-making. To overcome these limitations, we introduce a new method called Prediction-Augmented Generation (PAG). PAG can improve the generation quality and predictive accuracy of large language models in inference-driven tasks by integrating task-specific predictive models as external tools, enabling more structured and precise reasoning. Moreover, our method does not simply copy the inferences of a predictive model, but improves the inference results with knowledge from the large language model to create better predictions. We comprehensively evaluate our proposed method on diverse datasets for automatic diagnosis tasks requiring extensive domain knowledge and advanced reasoning.

Local Look-Ahead Guidance via Verifier-in-the-Loop for Automated Theorem Proving

Sara Rajaee, Kumar Pratik, Gabriele Cesa, Arash Behboodi

The most promising recent methods for AI reasoning require applying variants of reinforcement learning (RL) either on rolled out trajectories from the LLMs, even for the step-wise rewards, or large quantities of human-annotated trajectory data. The reliance on the rolled-out trajectory renders the compute cost and time prohibitively high. In particular, the correctness of a reasoning trajectory can typically only be judged at its completion, leading to sparse rewards in RL or requiring expensive synthetic data generation in expert iteration-like methods. In this work, we focus on the Automatic Theorem Proving (ATP) task and propose a novel verifier-in-the-loop design, which, unlike existing approaches that leverage feedback on the entire reasoning trajectory, employs an automated verifier to give intermediate feedback at each step of the reasoning process. Using Lean as the verifier, we empirically show that the step-by-step local verification produces a global improvement in the model's reasoning accuracy and efficiency.

Visualising Policy-Reward Interplay to Inform Zeroth-Order Preference Optimisation of Large Language Models

Alessio Galatolo, Zhenbang Dai, Meriem Beloucif, Katie Winkle

Fine-tuning LLMs with first-order methods like back-propagation is computationally intensive. Zeroth-Order (ZO) optimisation, using function evaluations instead of gradients, reduces memory usage but suffers from slow convergence in high-dimensional models. As a result, ZO research in LLMs has mostly focused on classification, overlooking more complex generative tasks. In this paper, we introduce ZOPrO, a novel ZO algorithm designed for *Preference Optimisation* in LLMs. We begin by analysing the interplay between policy and reward models during traditional (first-order) Preference Optimisation, uncovering patterns in their relative updates. Guided by these insights, we adapt Simultaneous Perturbation Stochastic Approximation (SPSA) with a targeted sampling strategy to accelerate convergence. Through experiments on summarisation, machine translation, and conversational assistants, we demonstrate that our method consistently enhances reward signals while achieving convergence times comparable to first-order methods. While it falls short of some state-of-the-art methods, our work is the first to apply Zeroth-Order methods to Preference Optimisation in LLMs, going beyond classification tasks and paving the way for a largely unexplored research direction.

VideoRAG: Retrieval-Augmented Generation over Video Corpus

Soyeong Jeong, Kangsan Kim, Jinheon Baek, Sung Ju Hwang

Retrieval-Augmented Generation (RAG) is a powerful strategy for improving the factual accuracy of models by retrieving external knowledge relevant to queries and incorporating it into the generation process. However, existing approaches primarily focus on text, with some recent advancements considering images, and they largely overlook videos, a rich source of multimodal knowledge capable of representing contextual details more effectively than any other modality. While very recent studies explore the use of videos in response generation, they either predefine query-associated videos without retrieval or convert videos into textual descriptions losing multimodal richness. To tackle these, we introduce VideoRAG, a framework that not only dynamically retrieves videos based on their relevance with queries and but also utilizes both visual and textual information. The operation of VideoRAG is powered by recent Large Video Language Models (LVLMs), which enable the direct processing of video content to represent it for retrieval and the seamless integration of retrieved videos jointly with queries for response generation. Also, inspired by that the context size of LVLMs may not be sufficient to process all frames in extremely long videos and not all frames are equally important, we introduce a video frame selection mechanism to extract the most informative subset of frames, along with a strategy to extract textual information from videos (as it can aid the understanding of video content) when their subtitles are not available. We experimentally validate the effectiveness of VideoRAG, showcasing that it is superior to relevant baselines.

Task Facet Learning: A Structured Approach To Prompt Optimization

Gurusha Juneja, Gautam Jajoo, Hua Li, Jian Jiao, Nagarajan Natarajan, Amit Sharma

Given a task in the form of a basic description and its training examples, prompt optimization is the problem of synthesizing the given information into a text prompt for a large language model. Humans solve this problem by also considering the different facets that define a task (e.g., counter-examples, explanations, analogies) and including them in the prompt. However, it is unclear whether existing algorithmic approaches, based on iteratively editing a given prompt or automatically selecting a few in-context examples, can cover the multiple facets required to solve a complex task. In this work, we view prompt optimization as that of learning multiple facets of a task from a set of training examples. We exploit structure in the prompt optimization problem and break down a prompt into loosely coupled semantic sections. The proposed algorithm, UniPrompt, (1) clusters the input space and uses clustered batches so that each batch likely corresponds to a different facet of the task, and (2) utilizes a feedback mechanism to propose adding, editing or deleting a section, which in turn is aggregated over a batch to capture generalizable facets. Empirical evaluation on multiple datasets and a real-world task shows that prompts generated using UniPrompt obtain higher accuracy than human-tuned prompts and those from state-of-the-art methods. In particular, our algorithm can generate long, complex prompts that existing methods are unable to generate.

Mosaic-IT: Cost-Free Compositional Data Synthesis for Instruction Tuning

Ming Li, Pei Chen, Chenguang Wang, Hongyu Zhao, Yijun Liang, YuPeng Hou, Fuxiao Liu, Tianyi Zhou

Finetuning large language models with a variety of instruction-response pairs has enhanced their capability to understand and follow instructions. Current instruction tuning primarily relies on teacher models or human intervention to generate and refine the instructions and responses for training, which are costly, non-sustainable, and may lack diversity. In this paper, we introduce Mosaic Instruction Tuning (Mosaic-IT), a human/model-free compositional data synthesis method that can efficiently create rich and diverse augmentations from existing instruction tuning data to enhance the LLMs. Mosaic-IT randomly concatenates multiple instruction data into one and trains the model to produce the corresponding responses with predefined higher-level meta-instructions to strengthen its multi-step instruction-following and format-following skills. Our extensive evaluations demonstrate a superior performance and training efficiency of Mosaic-IT, which achieves consistent performance improvements over various benchmarks and a reduction in training costs compared with the original instruction tuning.

Fuzzy Speculative Decoding for a Tunable Accuracy-Runtime Tradeoff

Maximilian Holsman, Yukun Huang, Bhuwan Dhingra

Speculative Decoding (SD) enforces strict distributional equivalence to the target model, limiting potential speed ups as distributions of near-equivalence achieve comparable outcomes in many cases. Furthermore, enforcing distributional equivalence means that users are unable to trade deviations from the target model distribution for further inference speed gains. To address these limitations, we introduce Fuzzy Speculative Decoding (FSD) - a decoding algorithm that generalizes SD by accepting candidate tokens purely based on the divergences between the target and draft model distributions. By allowing for controlled divergence from the target model, FSD enables users to flexibly trade generation quality for inference speed. Across several benchmarks, our method is able to achieve significant runtime improvements of over 5 tokens per second faster than SD at only an approximate 2% absolute reduction in benchmark accuracy. In many cases, FSD is even able to match SD benchmark accuracy at over 2 tokens per second faster, demonstrating that distributional equivalence is not necessary to maintain target model performance.

Statistical Deficiency for Task Inclusion Estimation

Loïc Fosse, Frederic Bechet, Benoit Favre, Géraldine Damnati, Gwénolé Lecorvé, Maxime DARRIN, Philippe Formont, Pablo Piantanida

Tasks are central in machine learning, as they are the most natural objects to assess the capabilities of current models. The trend is to build general models able to address any task. Even though transfer learning and multitask learning try to leverage the underlying task space, no well-founded tools are available to study its structure. This study proposes a theoretically grounded setup to define the notion of task and to compute the inclusion between two tasks from a statistical deficiency point of view. We propose a tractable proxy as information sufficiency to estimate the degree of inclusion between tasks, show its soundness on synthetic data, and use it to reconstruct empirically the classic NLP pipeline.

Dynamic Label Name Refinement for Few-Shot Dialogue Intent Classification

Gyutae Park, Ingeol Baek, Byeongjeong Kim, Joongbo Shin, Hwanhee Lee

Dialogue intent classification aims to identify the underlying purpose or intent of a user's input in a conversation. Current intent classification systems encounter considerable challenges, primarily due to the vast number of possible intents and the significant semantic overlap among similar intent classes. In this paper, we propose a novel approach to few-shot dialogue intent classification through in context learning, incorporating dynamic label refinement to address these challenges. Our method retrieves relevant examples for a test input from the training set and leverages a large language model to dynamically refine intent labels based on semantic understanding, ensuring that intents are clearly distinguishable from one another. Experimental results demonstrate that our approach effectively resolves confusion between semantically similar intents, resulting in significantly enhanced performance across multiple datasets compared to baselines. We also show that our method generates more interpretable intent labels, and has a better semantic coherence in capturing underlying user intents compared to baselines.

61-MAIN S³ - Semantic Signal Separation

Márton Kardos, Jan Kostkan, Kenneth Enevoldsen, Arnault-Quentin Vermillet, Kristoffer Nielbo, Roberta Rocca

Topic models are useful tools for discovering latent semantic structures in large textual corpora. Recent efforts have been oriented at incorporating contextual representations in topic modeling and have been shown to outperform classical topic models. These approaches are typically slow, volatile, and require heavy preprocessing for optimal results. We present Semantic Signal Separation (S^3), a theory-driven topic modeling approach in neural embedding spaces. S^3 conceptualizes topics as independent axes of semantic space and uncovers these by decomposing contextualized document embeddings using Independent Component Analysis. Our approach provides diverse and highly coherent topics, requires no preprocessing, and is demonstrated to be the fastest contextual topic model, being, on average, 4.5x faster than the runner-up BERTopic. We offer an implementation of S^3, and all contextual baselines, in the Turftopic Python package.

Doc-React: Multi-page Heterogeneous Document Question-answering

Junda Wu, Yu Xia, Tong Yu, Xiang Chen, Sai Sree Harsha, Akash V Maharaj, Ruiyi Zhang, Victor Bursztyn, Sungchul Kim, Ryan A. Rossi, Julian McAuley, Yunyao Li, Ritwik Sinha

Answering questions over multi-page, multimodal documents, including text and figures, is a critical challenge for applications that require answers to integrate information across multiple modalities and contextual dependencies. Existing methods, such as single-turn retrieval-augmented generation (RAG), struggle to retrieve fine-grained and contextually relevant information from large, heterogeneous documents, leading to suboptimal performance. Inspired by iterative frameworks like ReAct, which refine retrieval through feedback, we propose Doc-React, an adaptive iterative framework that balances information gain and uncertainty reduction at each step. Doc-React leverages InfoNCE-guided retrieval to approximate mutual information, enabling dynamic sub-query generation and refinement. A large language model (LLM) serves as both a judge and generator, providing structured feedback to iteratively improve retrieval. By combining mutual information optimization with entropy-aware selection, Doc-React systematically captures relevant multimodal content, achieving strong performance on complex QA tasks

Neural Topic Modeling with Large Language Models in the Loop

Xiaohao Yang, He Zhao, Weijie Xu, YUANYUAN QI, Jueqing Lu, Dinh Phung, Lan Du

Topic modeling is a fundamental task in natural language processing, allowing the discovery of latent thematic structures in text corpora. While Large Language Models (LLMs) have demonstrated promising capabilities in topic discovery, their direct application to topic modeling suffers from issues such as incomplete topic coverage, misalignment of topics, and inefficiency. To address these limitations, we propose LLM-ITL, a novel LLM-in-the-loop framework that integrates LLMs with Neural Topic Models (NTMs). In LLM-ITL, global topics and document representations are learned through the NTM. Meanwhile, an LLM refines these topics using an Optimal Transport (OT)-based alignment objective, where the refinement is dynamically adjusted based on the LLM's confidence in suggesting topical words for each set of input words. With the flexibility of being integrated into many existing NTMs, the proposed approach enhances the interpretability of topics while preserving the efficiency of NTMs in learning topics and document representations. Extensive experiments demonstrate that LLM-ITL helps NTMs significantly improve their topic interpretability while maintaining the quality of document representation. Our code and datasets are available at https://github.com/Xiaohao-Yang/LLM-ITL

Contrastive Perplexity for Controlled Generation: An Application in Detoxifying Large Language Models

Tassilo Klein, Moin Nabi

The generation of toxic content by large language models (LLMs) remains a critical challenge for the safe deployment of language technology. We propose a novel framework for implicit knowledge editing and controlled text generation by fine-tuning LLMs with a prototype-based contrastive perplexity objective. Central to our method is the construction of hard negatives—toxic outputs that are generated through adversarial paraphrasing to be semantically similar and model probability to their non-toxic counterparts. By training on these challenging and realistic pairs, our approach ensures robust and stable contrastive optimization. Experimental results in the domain of detoxification demonstrate that our method significantly reduces toxic generation while maintaining strong performance on downstream tasks such as commonsense reasoning and reading comprehension. Our findings highlight the effectiveness of exploiting hard negatives for attribute-aware fine-tuning.

HyperFM: Fact-Centric Multimodal Fusion for Link Prediction over Hyper-Relational Knowledge Graphs

Yuhuan Lu, Weijian Yu, Xin Jing, Dingqi Yang

With the ubiquity of hyper-relational facts in modern Knowledge Graphs (KGs), existing link prediction techniques mostly focus on learning the sophisticated relationships among multiple entities and relations contained in a fact, while ignoring the multimodal information, which often provides additional clues to boost link prediction performance. Nevertheless, traditional multimodel fusion approaches, which are mainly designed for triple facts under either entity-centric or relation-guided fusion schemes, fail to integrate the multimodal information with the rich context of the hyper-relational fact consisting of multiple entities and relations. Against this background, we propose **HyperFM**, a **Hyper**-relational **F**act-centric **M**ultimodal Fusion technique. It effectively captures the intricate interactions between different data modalities while accommodating the hyper-relational structure of the KG in a fact-centric manner via a customized Hypergraph Transformer. We evaluate HyperFM against a sizeable collection of baselines in link prediction tasks on two real-world KG datasets. Results show that HyperFM consistently achieves the best performance, yielding an average improvement of 6.0-6.8% over the best-performing baselines on the two datasets. Moreover, a series of ablation studies systematically validate our fact-centric fusion scheme.

SR-LLM: Rethinking the Structured Representation in Large Language Model

Jiahuan Zhang, Tianheng Wang, Ziyi Huang, Yulong Wu, HANQING WU, DongbaiChen, Linfeng Song, Yue Zhang, guozheng rao, Kaicheng Yu

Structured representations, exemplified by Abstract Meaning Representation(AMR), have long been pivotal in computational linguistics. However, their role remains ambiguous in the Large Language Models (LLMs) era. Initial attempts to integrate structured representation into LLMs via a zero-shot setting yielded inferior performance. We hypothesize that such a decline stems from the structure information being passed into LLMs in a code format unfamiliar to LLMs' training corpora. Consequently, we propose SR-LLM, an innovative framework with two settings to explore a superior way of integrating structured representation with LLMs from training-free and training-dependent perspectives. The former integrates structural information through natural language descriptions in LLM prompts, whereas its counterpart augments the model's inference capability through fine-tuning on linguistically described structured representations. Performance improvements were observed in widely downstream datasets, with particularly notable gains of 3.17% and 12.38% in PAWS. To the best of our knowledge, this work represents the pioneering demonstration that leveraging structural representations can substantially enhance LLMs' inference capability. We hope that our work sheds light and encourages future research to enhance the reasoning and interoperability of LLMs by structure data.

Contrastive Prompting Enhances Sentence Embeddings in LLMs through Inference-Time Steering

Zifeng Cheng, Zhonghui Wang, Yuchen Fu, Zhiwei Jiang, Yafeng Yin, Cong Wang, Qing Gu

Extracting sentence embeddings from large language models (LLMs) is a practical direction, as it requires neither additional data nor fine-tuning. Previous studies usually focus on prompt engineering to guide LLMs to encode the core semantic information of the sentence into the embedding of the last token. However, the last token in these methods still encodes an excess of non-essential information, such as stop words, limiting its encoding capacity. To this end, we propose a Contrastive Prompting (CP) technique that introduces an extra auxiliary prompt to elicit better sentence embedding. By contrasting with the auxiliary prompt, CP can steer existing prompts to encode the core semantics of the sentence, rather than non-essential information. CP is a plug-and-play inference-time intervention method that can be combined with various prompt-based methods. Extensive experiments on Semantic Textual Similarity (STS) tasks and downstream classification tasks demonstrate that our method can improve the performance of existing prompt-based methods across different LLMs.

Large Margin Representation Learning for Robust Cross-lingual Named Entity Recognition

Guangcheng Zhu, Ruixuan Xiao, Zhen Zhu, Gengyu Lyu, Junbo Zhao, Haobo Wang

Cross-lingual named entity recognition (NER) aims to build an NER model that generalizes to the low-resource target language with labeled data from the high-resource source language. Current state-of-the-art methods typically combine self-training mechanism with contrastive learning paradigm, in order to develop discriminative entity clusters for cross-lingual adaptation. Despite the promise, we identify that these methods neglect two key problems: distribution skewness and pseudo-label bias, leading to indistinguishable entity clusters with small margins. To this end, we propose a novel framework, MARAL, which optimizes an adaptively reweighted contrastive loss to handle the class skewness and theoretically guarantees the optimal feature arrangement with maximum margin. To further mitigate the adverse effects of unreliable pseudo-labels, MARAL integrates a progressive cross-lingual adaptation strategy, which first selects reliable samples as anchors and then refines the remaining unreliable ones. Extensive experiments demonstrate that MARAL significantly outperforms the current state-of-the-art methods on multiple benchmarks, e.g., +2.04% on the challenging MultiCoNER dataset.

PTQ1.61: Push the Real Limit of Extremely Low-Bit Post-Training Quantization Methods for Large Language Models

Jiaqi Zhao, Miao Zhang, Ming Wang, Yuzhang Shang, Kaihao Zhang, Weili Guan, Yaowei Wang, Min Zhang

Large Language Models (LLMs) suffer severe performance degradation when facing extremely low-bit (sub 2-bit) quantization. Several existing sub 2-bit post-training quantization (PTQ) methods utilize a mix-precision scheme by leveraging an unstructured fine-grained mask to explicitly distinguish salient weights, while which introduces an extra 1-bit or more per weight. To explore the real limit of PTQ, we propose an extremely low-bit PTQ method called PTQ1.61, which enables weight quantization to 1.61-bit for the first time. Specifically, we first introduce a one-dimensional structured mask with negligibly additional 0.0002-bit per weight based on input activations from the perspective of reducing the upper bound of quantization error to allocate corresponding salient weight channels to 4-bit. For non-salient channels binarization, an efficient block-wise scaling factors optimization framework is then presented to take implicit row-wise correlations and angular biases into account. Different from prior works that concentrate on adjusting quantization methodologies, we further propose a novel paradigm called quantization preprocessing, where we argue that transforming the weight distribution of the pretrained model before quantization can alleviate the difficulty in per-channel extremely low-bit PTQ. Extensive experiments indicate our PTQ1.61 achieves state-of-the-art performance in extremely low-bit quantization. Codes are available at https://github.com/zjq0455/PTQ1.61.

GradOT: Training-free Gradient-persevering Offsite-tuning for Large Language Models

Kai Yao, Zhaorui Tan, Penglei Gao, Lichun Li, Kaixin Wu, Yinggui Wang, Yuan Zhao, Yixin Ji, Jianke Zhu, Wei Wang

The rapid growth of large language models (LLMs) with traditional centralized fine-tuning emerges as a key technique for adapting these models to domain-specific challenges, yielding privacy risks for both model and data owners. One promising solution, called offsite-tuning (OT), is proposed to address these challenges, where a weaker emulator is compressed from the original model and further fine-tuned with adapter to enhance privacy. However, the existing OT-based methods require high computational costs and lack theoretical analysis. This paper introduces a novel OT approach based on gradient-preserving compression. By analyzing the OT problem through the lens of optimization, we propose a method that selectively applies compression techniques such as rank compression and channel pruning, preserving the gradients of fine-tuned adapters while ensuring privacy. Extensive experiments demonstrate that our approach surpasses existing OT methods, both in terms of privacy protection and model performance. Our method provides a theoretical foundation for OT and offers a practical, training-free solution for offsite-tuning of large-scale LLMs.

Modality-Aware Neuron Pruning for Unlearning in Multimodal Large Language Models

Zheyuan Liu, Guangyao Dou, Xiangchi Yuan, Chunhui Zhang, Zhaoxuan Tan, Meng Jiang

Generative models such as Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs) trained on massive datasets can lead them to memorize and inadvertently reveal sensitive information, raising ethical and privacy concerns. While some prior works have explored this issue in the context of LLMs, it presents a unique challenge for MLLMs due to the entangled nature of knowledge across modalities, making comprehensive unlearning more difficult. To address this challenge, we propose Modality Aware Neuron Unlearning (MANU), a novel unlearning framework for MLLMs designed to selectively clip neurons based on their relative importance to the targeted forget data, curated for different modalities. Specifically, MANU consists of two stages: important neuron selection and selective pruning. The first stage identifies and collects the most influential neurons across modalities relative to the targeted forget knowledge, while the second stage is dedicated to pruning those selected neurons. MANU effectively isolates and removes the neurons that contribute most to the forget data within each modality, while preserving the integrity of retained knowledge. Our experiments conducted across various MLLM architectures illustrate that MANU can achieve a more balanced and comprehensive unlearning in each modality without largely affecting the overall model utility.

ReLearn: Unlearning via Learning for Large Language Models

Haoming Xu, Ningyuan Zhao, Liming Yang, Sendong Zhao, Shumin Deng, Mengru Wang, Bryan Hooi, Nay Oo, Huajun Chen, Ningyu Zhang

Current unlearning methods for large language models usually rely on reverse optimization to reduce target token probabilities. However, this paradigm disrupts the subsequent tokens prediction, degrading model performance and linguistic coherence. Moreover, existing evaluation metrics overemphasize contextual forgetting while inadequately assessing response fluency and relevance. To address these challenges, we propose ReLearn, a data augmentation and fine-tuning pipeline for effective unlearning, along with a comprehensive evaluation framework. This framework introduces Knowledge Forgetting Ratio (KFR) and Knowledge Retention Ratio (KRR) to measure knowledge-level preservation, and Linguistic Score (LS) to evaluate generation quality. Our experiments show that ReLearn successfully achieves targeted forgetting while preserving high-quality outputs. Through mechanistic analysis, we further demonstrate how reverse optimization disrupts coherent text generation, while ReLearn preserves this essential capability.

CxGGEC: Construction-Guided Grammatical Error Correction

Yayu Cao, Tianxiang Wang, Lvxiaowei Xu, Zhenyao Wang, Ming Cai

The grammatical error correction (GEC) task aims to detect and correct grammatical errors in text to enhance its accuracy and readability. Current GEC methods primarily rely on grammatical labels for syntactic information, often overlooking the inherent usage patterns of language. In this work, we explore the potential of construction grammar (CxG) to improve GEC by leveraging constructions to capture underlying language patterns and guide corrections. We first establish a comprehensive construction inventory from corpora. Next, we introduce a construction prediction model that identifies potential constructions in ungrammatical sentences using a noise-tolerant language model. Finally, we train a CxGGEC model on construction-masked parallel data, which performs GEC by decoding construction tokens into their original forms and correcting erroneous tokens. Extensive experiments on English and Chinese GEC benchmarks demonstrate the effectiveness of our approach.

Instance-Selection-Inspired Undersampling Strategies for Bias Reduction in Small and Large Language Models for Binary Text Classification

Guilherme Fonseca, Washington Cunha, Gabriel Prenassi, Marcos André Gonçalves, Leonardo Chaves Dutra da Rocha

Skewness in imbalanced datasets affects Automatic Text Classification (ATC), leading to classifier bias toward the majority classes. This work examines undersampling methods to mitigate such bias in Small and Large Language Model (SLMs and LLMs) classifiers. Based on the limitations found in existing solutions, we propose two novel undersampling methods inspired by state-of-the-art Instance Selection techniques, relying on calibrated confidences and semantic difficulty estimates. We compare them against 19 baselines across 13 datasets, evaluating: (i) effectiveness, (ii) class imbalance bias, (iii) efficiency, (iv) scalability, and (v) consistency. Results show our methods uniquely reduce classifier bias (up to 56%) across all datasets without effectiveness loss while improving efficiency (1.6x speedup), scalability and reducing carbon emissions (up to 50%).

SynWorld: Virtual Scenario Synthesis for Agentic Action Knowledge Refinement

Runnan Fang, Xiaobin Wang, Yuan Liang, Shuofei Qiao, Jialong Wu, Zekun Xi, Ningyu Zhang, Yong Jiang, Pengjun Xie, Fei Huang, Huajun Chen

In the interaction between agents and their environments, agents expand their capabilities by planning and executing actions. However, LLM-based agents face substantial challenges when deployed in novel environments or required to navigate unconventional action spaces. To empower agents to autonomously explore environments, optimize workflows, and enhance their understanding of actions, we propose SynWorld, a framework that allows agents to synthesize possible scenarios with multi-step action invocation within the action space and perform Monte Carlo Tree Search (MCTS) exploration to effectively refine their action knowledge in the current environment. Our experiments demonstrate that SynWorld is an effective and general approach to learning action knowledge in new environments.

A Strategic Coordination Framework of Small LMs Matches Large LMs in Data Synthesis

Xin Gao, Qizhi Pei, Zinan Tang, Yu Li, Honglin Lin, Jiang Wu, Lijun Wu, Conghui He

While data synthesis and distillation are promising strategies to enhance small language models, current approaches heavily rely on Large Language Models (LLMs), which suffer from high computational costs, environmental inefficiency, and potential biases inherited from monolithic architectures. In contrast, smaller LMs are more accessible and sustainable, but their individual capabilities often fall short in generating high-quality, diverse, and reliable data. Inspired by collaborative human processes (e.g., peer review), we propose a multiple small LMs involved framework, GRA, that aggregates specialized roles across small LMs to iterative refinement and quality control typically achieved by a single large LM. In this collaborative framework, multiple small LMs assume distinct roles—Generator, Reviewer, and Adjudicator—to simulate a peer-review-inspired data synthesis pipeline. The Generator proposes initial data samples, the Reviewer critiques their quality and diversity, and the Adjudicator resolves conflicts to finalize the output. By decomposing the synthesis process into specialized sub-tasks, collaborative small LMs can achieve data-level parity with distillation from large LMs. Through experiments across multiple benchmarks, we demonstrate that GRA-produced data matches or exceeds the quality of single large LM outputs, e.g., Qwen-2.5-72B-Instruct. Our results challenge the necessity of monolithic large models for high-quality data synthesis, advocating instead for strategic coordination of smaller agents.

EpMAN: Episodic Memory AttentioN for Generalizing to Longer Contexts

Subhajit Chaudhury, Payel Das, Sarathkrishna Swaminathan, Georgios Kollias, Elliot Nelson, Khushbu Pahwa, Tejaswini Pedapati, Igor Melnyk, Matthew Riemer

Recent advances in Large Language Models (LLMs) have yielded impressive successes on many language tasks. However, efficient processing of long contexts using LLMs remains a significant challenge. We introduce **EpMAN** -- a method for processing long contexts in an episodic memory module while holistically attending to semantically-relevant context chunks. Output from episodic attention is then used to reweigh the decoder's self-attention to the stored KV cache of the context during training and generation. When an LLM decoder is trained using **EpMAN**, its performance on multiple challenging single-hop long-context recall and question-answering benchmarks is found to be stronger and more robust across the range from 16k to 256k tokens than baseline decoders trained with self-attention, and popular retrieval-augmented generation frameworks.

Agentic Knowledgeable Self-awareness

Shuofei Qiao, Zhisong Qiu, Baochang Ren, Xiaobin Wang, Xiangyuan Ru, Ningyu Zhang, Xiang Chen, Yong Jiang, Pengjun Xie, Fei Huang, Huajun Chen

Large Language Models (LLMs) have achieved considerable performance across various agentic planning tasks. However, traditional approaches adopt a ``flood irrigation'' methodology that indiscriminately injects gold trajectories, external feedback, and domain knowledge into agent models. This practice overlooks the fundamental human cognitive principle of self-awareness - the ability to dynamically assess situational demands and strategically employ resources during decision-making. We propose \textbf{Agentic Knowledgeable Self-awareness} to address this gap, a novel paradigm enabling LLM-based agents to autonomously regulate knowledge utilization. Specifically, we propose \texttt{KnowSelf}, a data-centric approach that applies agents with \texttt{know}ledgeable \texttt{self}-awareness like humans. Concretely, we devise a heuristic situation judgement criterion to mark special tokens on the agent's self-explored trajectories for collecting training data. Through a two-stage training process, the agent model can switch between different situations by generating specific special tokens, achieving optimal planning effects with minimal costs. Our experiments demonstrate that can outperform various strong baselines on different tasks and models with minimal use of external knowledge.

Multi-task Adversarial Attacks against Black-box Model with Few-shot Queries

Wenqiang Wang, Yan XIAO, Hao Lin, Yangshijie Zhang, Xiaochun Cao

Current multi-task adversarial text attacks rely on abundant access to shared internal features and numerous queries, often limited to a single task type. As a result, these attacks are less effective against practical scenarios involving black-box feedback APIs, limited queries, or multiple task types. To bridge this gap, we propose \textbf{C}luster and \textbf{E}nsemble \textbf{M}util-task Text Adversarial \textbf{A}ttack (\textbf{CEMA}), an effective black-box attack that exploits the transferability of adversarial texts across different tasks. CEMA simplifies complex multi-task scenarios by using a \textit{deep-level substitute model} trained in a \textit{plug-and-play} manner for text classification, enabling attacks without mimicking the victim model. This approach requires only a few queries for training, converting multi-task attacks into classification attacks and allowing attacks across various tasks. CEMA generates multiple adversarial candidates using different text classification methods and selects the one that most effectively attacks substitute models. In experiments involving multi-task models with two, three, or six tasks—spanning classification, translation, summarization, and text-to-image generation—CEMA demonstrates significant attack success with as few as 100 queries. Furthermore, CEMA can target commercial APIs (e.g., Baidu and Google Translate), large language models (e.g., ChatGPT 4o), and image-generation models (e.g., Stable Diffusion V2), showcasing its versatility and effectiveness in real-world applications.

Flexora: Flexible Low-Rank Adaptation for Large Language Models

Chenxing Wei, Yao Shu, Ying Tiffany He, Fei Yu

Large language models (LLMs) have revolutionized artificial intelligence, but their performance on specific tasks is often limited by knowledge boundaries. While fine-tuning techniques like low-rank adaptation (LoRA) aim to address this, they can suffer from overfitting. We propose flexible low-rank adaptation (Flexora), a novel method that automatically selects the most critical layers for fine-tuning to optimize performance across diverse downstream tasks. Flexora formulates layer selection as a hyperparameter optimization problem, employs unrolled differentiation for efficient solving, and identifies the most impactful layers based on optimized hyperparameters. Extensive experiments across various pre-trained models and natural language tasks demonstrate that Flexora consistently outperforms existing baselines. We provide theoretical insights and comprehensive ablation studies to elucidate the effectiveness of Flexora. Therefore, Flexora offers a robust solution to enhance LoRA fine-tuning for LLMs, potentially advancing the field of adaptive language model optimization.

SAKE: Steering Activations for Knowledge Editing

Marco Scialanga, Thibault Laugel, Vincent Grari, Marcin Detyniecki

As Large Langue Models have been shown to memorize real-world facts, the need to update this knowledge in a controlled and efficient manner arises. Designed with these constraints in mind, Knowledge Editing (KE) approaches propose to alter specific facts in pretrained models. However, they have been shown to suffer from several limitations, including their lack of contextual robustness and their failure to generalize to logical implications related to the fact. To overcome these issues, we propose SAKE, a steering activation method that models a fact to be edited as a distribution rather than a single prompt. Leveraging Optimal Transport, SAKE alters the LLM behavior over a whole fact-related distribution, defined as paraphrases and logical implications. Several numerical experiments demonstrate the effectiveness of this method: SAKE is thus able to perform more robust edits than its existing counterparts.

Fixing Distribution Shifts of LLM Self-Critique via On-Policy Self-Play Training

Rong Bao, Donglei Yu, Kai Fan, Minpeng Liao

Self-critique mechanisms significantly improve the performance of language models in complex reasoning tasks by giving them the ability to correct errors, conduct induction and deduction, and switch thinking insights. However, synthetic data methods often require human-introduced errors or sampling of the model's reasoning results from the previous moment, and the current output distribution of the model cannot be obtained, makes the data for critique and reasoning face the problem of distribution shifts. In this work, we propose an on-policy reinforcement learning framework to synchronize the reasoning and critique capabilities of language models. To alleviate reward hacking caused by outcome-based supervision, we design a deliberate reward framework for different purposes. The reward framework not only supervises the model reasoning process based on the results, but also uses Monte Carlo sampling to give appropriate rewards to the critique content according to the success rate of the model's correction after critique. In addition, we introduce a rule-based reward function to impose penalties on the model when it generates hallucinatory critiques. When our approach is applied to the DeepSeek-Math-7B-Base and Qwen2.5-7B-Base models, model performance improves 5.40 and 3.66 points, respectively, compared to the best baseline approach. This validates the significant advantages of our method in improving model's reasoning and self-critique capability. Code will be made available at https://github.com/rbao2018/SCOP

HAF-RM: A Hybrid Alignment Framework for Reward Model Training

Shujun Liu, Xiaoyu Shen, Yuhang Lai, Siyuan Wang, ShengbinYue, Zengfeng Huang, Xuanjing Huang, zhongyu wei

The reward model has become increasingly important in alignment, assessment, and data construction for large language models (LLMs). Most existing researchers focus on enhancing reward models through data improvements, following the conventional training framework for reward models that directly optimizes the predicted rewards. In this paper, we propose a hybrid alignment framework **HAF-RM** for reward model training by introducing an additional constraint on token-level policy probabilities in addition to the reward score. It can simultaneously supervise the internal preference model at the token level and optimize the mapping layer of the reward model at the sequence level. Experiment results on five datasets sufficiently show the validity and effectiveness of our proposed hybrid framework for training a high-quality reward model. By decoupling the reward modeling procedure and incorporating hybrid supervision, our **HAF-RM** framework offers a principled and effective approach to enhancing the performance and alignment of reward models, a critical component in the responsible development of powerful language models. We release our code at https://haf-rm.github.io.

Segment-Based Attention Masking for GPTs

Shahar Katz, Liran Ringel, Yaniv Romano, Lior Wolf

Causal masking is a fundamental component in Generative Pre-Trained Transformer (GPT) models, playing a crucial role during training. Although GPTs can process the entire user prompt at once, the causal masking is applied to all input tokens step-by-step, mimicking the generation process. This imposes an unnecessary constraint during the initial "prefill" phase when the model processes the input prompt and generates the internal representations before producing any output tokens. In this work, attention is masked based on the known block structure at the prefill phase, followed by the conventional token-by-token autoregressive process after that. For example, in a typical chat prompt, the system prompt is treated as one block, and the user prompt as the next one. Each of these is treated as a unit for the purpose of masking, such that the first tokens in each block can access the subsequent tokens in a non-causal manner. Then, the model answer is generated in the conventional causal manner. The Segment-by-Segment scheme entails no additional computational overhead. When integrated using a lightweight fine-tuning into already trained models such as Llama and Qwen, MAS quickly increases models' performances.

A Modular Approach for Clinical SLMs Driven by Synthetic Data with Pre-Instruction Tuning, Model Merging, and Clinical-Tasks Alignment

Jean-Philippe Corbeil, Amin Dada, Jean-Michel Attendu, Asma Ben Abacha, Alessandro Sordoni, Lucas Caccia, Francois Beaulieu, Thomas Lin, Jens Kleesiek, Paul Vozila

High computation costs and latency of large language models such as GPT-4 have limited their deployment in clinical settings. Small language models (SLMs) offer a cost-effective alternative, but their limited capacity requires biomedical domain adaptation, which remains challenging. An additional bottleneck is the unavailability and high sensitivity of clinical data. To address these challenges, we propose a novel framework for adapting SLMs into high-performing clinical models. We introduce the MediPhi collection of 3.8B-parameter SLMs developed with our novel framework: pre-instruction tuning of experts on relevant medical and clinical corpora (PMC, Medical Guideline, MedWiki, etc.), model merging, and clinical-tasks alignment. To cover most clinical tasks, we extended the CLUE benchmark to CLUE+, doubling its size. Our expert models deliver relative improvements on this benchmark over the base model without any task-specific fine-tuning: 64.3% on medical entities, 49.5% on radiology reports, and 44% on ICD-10 coding (outperforming GPT-4-0125 by 14%). We unify the expert models into MediPhi via model merging, preserving gains across benchmarks. Furthermore, we built the MediFlow collection, a synthetic dataset of 2.5 million high-quality instructions on 14 medical NLP tasks, 98 fine-grained document types, and JSON format support. Alignment of MediPhi using supervised fine-tuning and direct preference optimization achieves further gains of 18.9% on average.

Impartial Multi-task Representation Learning via Variance-invariant Probabilistic Decoding

Dou Hu, Lingwei Wei, Wei Zhou, Songlin Hu

Multi-task learning (MTL) enhances efficiency by sharing representations across tasks, but task dissimilarities often cause partial learning, where some tasks dominate while others are neglected. Existing methods mainly focus on balancing loss or gradients but fail to fundamentally address this issue due to the representation discrepancy in latent space. In this paper, we propose variance-invariant probabilistic decoding for multi-task learning (VIP-MTL), a framework that ensures impartial learning by harmonizing representation spaces across tasks. VIP-MTL decodes shared representations into task-specific probabilistic distributions and applies variance normalization to constrain these distributions to a consistent scale. Experiments on two language benchmarks show that VIP-MTL outperforms 12 representative methods under the same multi-task settings, especially in heterogeneous task combinations and data-constrained scenarios. Further analysis shows that VIP-MTL is robust to sampling distributions, efficient on optimization process, and scale-invariant to task losses. Additionally, the learned task-specific representations are more informative, enhancing the language understanding abilities of pre-trained language models under the multi-task paradigm.

AutoMixAlign: Adaptive Data Mixing for Multi-Task Preference Optimization in LLMs

Nicholas E. Corrado, Julian Katz-Samuels, Adithya M Devraj, Hyokun Yun, Chao Zhang, Yi Xu, Yi Pan, Bing Yin, Trishul Chilimbi

When aligning large language models (LLMs), their performance across various tasks (such as being helpful, harmless, and honest) is heavily influenced by the composition of the training data. However, it is difficult to determine what mixture of data should be used to produce a model with strong performance across all tasks. Existing approaches rely on large ablation studies, heuristics, or human intuition, though these can be prohibitively expensive and suboptimal. We study this problem in the context of preference optimization via DPO and propose a novel and theoretically justified algorithm, AutoMixAlign (AMA), that adaptively mixes datasets during LLM training to balance performance across multiple tasks. AMA first trains \textit{specialist models} for each task to determine losses that corresponding to strong task performance. Next, AMA trains a generalist model using a novel minimax optimization that prioritizes tasks for which generalist model losses are furthest from specialist model losses. We introduce two algorithms to optimize this problem: (1) AMA-R adaptively reweights the objective to prioritize tasks, and (2) AMA-S adaptively adjusts how much data is sampled from each task to prioritize tasks. Both algorithms achieve a convergence rate of O(1/\sqrt{T}) in the convex case. AMA-R's convergence result immediately follows from Sagawa et. al, 2019, and we provide a convergence proof for AMA-S using techniques from online learning such as EXP3 (Auer et. al, 2002). We evaluate AMA on several multitask alignment setups, and observe that AMA outperforms the standard alignment approach which simply optimizes the total loss across all tasks and also outperforms model-merging methods.

A Mutual Information Perspective on Knowledge Graph Embedding

Jiang Li, Xiangdong Su, Zehua Duo, Tian Lan, Xiaotao Guo, Guanglai Gao

Knowledge graph embedding techniques have emerged as a critical approach for addressing the issue of missing relations in knowledge graphs. However, existing methods often suffer from limitations, including high intra-group similarity, loss of semantic information, and insufficient inference capability, particularly in complex relation patterns such as 1-N and N-1 relations. To address these challenges, we introduce a novel KGE framework that leverages mutual information maximization to improve the semantic representation of entities and relations. By maximizing the mutual information between different components of triples, such as \((h, r)\) and \(t\), or \((r, t)\) and \(h\), the proposed method improves the model's ability to preserve semantic dependencies while maintaining the relational structure of the knowledge graph. Extensive experiments on benchmark datasets demonstrate the effectiveness of our approach, with consistent performance improvements across various baseline models. Additionally, visualization analyses and case studies demonstrate the improved ability of the MI framework to capture complex relation patterns.

Detecting Sockpuppetry on Wikipedia Using Meta-Learning

Christine de Kock, Luc Raszewski

Malicious sockpuppet detection on Wikipedia is critical to preserving access to reliable information on the internet and preventing the spread of disinformation. Prior machine learning approaches rely on stylistic and meta-data features, but do not prioritise adaptability to author-specific behaviours. As a result, they struggle to effectively model the behaviour of specific sockpuppet-groups, especially when text data is limited. To address this, we propose the application of meta-learning, a machine learning technique designed to improve performance in data-scarce settings by training models across multiple tasks. Meta-learning optimises a model for rapid adaptation to the writing style of a new sockpuppet-group. Our results show that meta-learning significantly enhances the precision of predictions compared to pre-trained models, marking an advancement in combating sockpuppetry on open editing platforms. We release an updated dataset of sockpuppet investigations to foster future research in both sockpuppetry and meta-learning fields.

S2WTM: Spherical Sliced-Wasserstein Autoencoder for Topic Modeling

Suman Adhya, Debarshi Kumar Sanyal

Modeling latent representations in a hyperspherical space has proven effective for capturing directional similarities in high-dimensional text data, benefiting topic modeling. Variational autoencoder-based neural topic models (VAE-NTMs) commonly adopt the von Mises-Fisher prior to encode hyperspherical structure. However, VAE-NTMs often suffer from posterior collapse, where the KL divergence term in the objective function highly diminishes, leading to ineffective latent representations. To mitigate this issue while modeling hyperspherical structure in the latent space, we propose the Spherical Sliced Wasserstein Autoencoder for Topic Modeling (S2WTM). S2WTM employs a prior distribution supported on the unit hypersphere and leverages the Spherical Sliced-Wasserstein distance to align the aggregated posterior distribution with the prior. Experimental results demonstrate that S2WTM outperforms state-of-the-art topic models, generating more coherent and diverse topics while improving performance on downstream tasks.

Don't Half-listen: Capturing Key-part Information in Continual Instruction Tuning

Yongquan He, Wenyuan Zhang, Xuancheng Huang, peng zhang, Lingxun Meng, Xiang Zhou, Ke Zeng, Xunliang Cai

Instruction tuning for large language models (LLMs) can drive them to produce results consistent with human goals in specific downstream tasks. However, the process of continual instruction tuning (CIT) for LLMs may bring about the catastrophic forgetting (CF) problem, where previously learned abilities are degraded. Recent methods try to alleviate the CF problem by modifying models or replaying data, which may only remember the surface-level pattern of instructions and get confused on held-out tasks. In this paper, we propose a novel continual instruction tuning method based on Key-part Information Gain (KPIG). Our method computes the information gain on masked parts to dynamically replay data and refine the training objective, which enables LLMs to capture task-aware information relevant to the correct response and alleviate overfitting to general descriptions in instructions. In addition, we propose two metrics, P-score and V-score, to measure the generalization and instruction-following abilities of LLMs. Experiments demonstrate our method achieves superior performance on both seen and held-out tasks.

Graph of Records: Boosting Retrieval Augmented Generation for Long-context Summarization with Graphs

Haozhen Zhang, Tao Feng, Jiaxuan You

Retrieval-augmented generation (RAG) has revitalized Large Language Models (LLMs) by injecting non-parametric factual knowledge. Compared with long-context LLMs, RAG is considered an effective summarization tool in a more concise and lightweight manner, which can interact with LLMs multiple times using diverse queries to get comprehensive responses. However, the LLM-generated historical responses, which contain potentially insightful information, are largely neglected and discarded by existing approaches, leading to suboptimal results. In this paper, we propose \textit{graph of records} (\textbf{GoR}), which leverages historical responses generated by LLMs to enhance RAG for long-context global summarization. Inspired by the \textit{retrieve-then-generate} paradigm of RAG, we construct a graph by establishing an edge between the retrieved text chunks and the corresponding LLM-generated response. To further uncover the intricate correlations between them, GoR features a \textit{graph neural network} and an elaborately designed \textit{BERTScore}-based objective for self-supervised model training, enabling seamless supervision signal backpropagation between reference summaries and node embeddings. We comprehensively compare GoR with 12 baselines across four long-context summarization datasets, and the results indicate that our proposed method reaches the best performance (\textit{e.g.}, 15%, 8%, and 19% improvement over retrievers w.r.t. Rouge-L, Rouge-1, and Rouge-2 on the WCEP dataset). Extensive experiments further demonstrate the effectiveness of GoR.

Better Embeddings with Coupled Adam

Felix Stollenwerk, Tobias Stollenwerk

Despite their remarkable capabilities, LLMs learn word representations that exhibit the undesirable yet poorly understood feature of anisotropy. In this paper, we argue that the second moment in Adam is a cause of anisotropic embeddings, and suggest a modified optimizer called Coupled Adam to mitigate the problem. Our experiments demonstrate that Coupled Adam significantly improves the quality of embeddings, while also leading to better upstream and downstream performance on large enough datasets.

Bone Soups: A Seek-and-Soup Model Merging Approach for Controllable Multi-Objective Generation

Guofu Xie, Xiao Zhang, Ting Yao, Yunsheng Shi

User information needs are often highly diverse and varied. A key challenge in current research is how to achieve controllable multi-objective generation while enabling rapid adaptation to accommodate diverse user demands during test time. Existing solutions, such as Rewarded Soup, focus on merging language models individually tuned on single objectives. While easy to implement and widely used, these approaches face limitations in achieving optimal performance due to their disregard for the impacts of competing objectives on model tuning. To address this issue, we propose **Bone Soup**, a novel model merging approach that first seeks a series of back**bone** models by considering the impacts of multiple objectives and then makes the **soup** (i.e., merge the backbone models). Specifically, Bone Soup begins by training multiple backbone models for different objectives using multi-objective reinforcement learning. Each backbone model is guided by a combination of backbone reward signals. To ensure that these models are optimal for the Pareto front, the backbone rewards are crafted by combining standard reward functions into basis vectors, which can then be modified through a rule-based construction method. Bone Soup leverages a symmetric circulant matrix mapping to generate the merging coefficients, which are used to merge the backbone models according to user preferences. Extensive experimental results demonstrate that Bone Soup exhibits strong controllability and Pareto optimality in controllable multi-objective generation, providing a more effective and efficient approach to addressing diverse user needs at test time.

Recurrent Knowledge Localization and Fusion for Language Model Continual Learning

Yujie Feng, Xujia Wang, ZEXIN LU, FuShenghong, Guangyuan SHI, Yongxin Xu, Yasha Wang, Philip S. Yu, Xu Chu, Xiao-Ming Wu

Continual learning (CL) is crucial for deploying large language models (LLMs) in dynamic real-world environments without costly retraining. While recent model ensemble and model merging methods guided by parameter importance have gained popularity, they often struggle to balance knowledge transfer and forgetting, mainly due to the reliance on static importance estimates during sequential training. In this paper, we present Recurrent-KIF, a novel CL framework for Recurrent Knowledge Identification and Fusion, which enables dynamic estimation of parameter importance distributions to enhance knowledge transfer. Inspired by human continual learning, Recurrent-KIF employs an inner loop that rapidly adapts to new tasks while identifying important parameters, coupled with an outer loop that globally manages the fusion of new and historical knowledge through redundant knowledge pruning and key knowledge merging. These inner-outer loops iteratively perform multiple rounds of fusion, allowing Recurrent-KIF to leverage intermediate training information and adaptively adjust fusion strategies based on evolving importance distributions. Extensive experiments on two CL benchmarks with various model sizes (from 770M to 13B) demonstrate that Recurrent-KIF effectively mitigates catastrophic forgetting and enhances knowledge transfer.

MMDEND: Dendrite-Inspired Multi-Branch Multi-Compartment Parallel Spiking Neuron for Sequence Modeling

Kexin Wang, Yuhong Chou, Di Shang, Shijie Mei, Jiahong Zhang, Yanbin Huang, Man Yao, Bo XU, Guoqi Li

Vanilla spiking neurons are simplified from complex biological neurons with dendrites, soma, and synapses, into single somatic compartments. Due to limitations in performance and training efficiency, vanilla spiking neurons face significant challenges in modeling long sequences. In terms of performance, the oversimplified dynamics of spiking neurons omit long-term temporal dependencies. Additionally, the long-tail membrane potential distribution and binary activation discretization errors further limit their capacity to model long sequences. In terms of efficiency, the serial mechanism of spiking neurons leads to excessively long training times for long sequences. Though parallel spiking neurons are an efficient solution, their number of parameters is often tied to the hidden dimension or sequence length, which makes current parallel neurons unsuitable for large architectures. To address these issues, we propose **MMDEND**: a Multi-Branch Multi-Compartment Parallel Spiking Dendritic Neuron. Its proportion-adjustable multi-branch, multi-compartment structure enables long-term temporal dependencies. Additionally, we introduce a Scaling-Shifting Integer Firing (SSF) mechanism that fits the long-tail membrane potential distribution, retains efficiency, and mitigates discretization errors. Compared with parallel neurons, MMDEND achieves better long-sequence modeling capability with fewer parameters and lower energy consumption. Visualization also confirms that the SSF mechanism effectively fits long-tail distributions.

Language Models, Graph Searching, and Supervision Adulteration: When More Supervision is Less and How to Make More More

Arvid Frydenlund

This work concerns the path-star task, a minimal example of searching over a graph. The graph, G, is star-shaped with D arms radiating from a start node, s. A language model (LM) is given G, s, and a target node, t, which ends one of the arms and is tasked with generating the arm containing t. The minimal nature of this task means only a single choice needs to be made: which of the arms contains? Decoder-only LMs fail to solve this elementary task above 1/D chance due to a learned shortcut that absorbs training supervision. We show how this pathology is caused by excess supervision and present a series of solutions demonstrating that the task is solvable via decoder-only LMs. We find that the task's minimal nature causes its difficulty, as it prevents task decomposition. Our solutions provide insight into the pathology and its implications for LMs trained via next-token prediction.

MAPoRL: Multi-Agent Post-Co-Training for Collaborative Large Language Models with Reinforcement Learning

Chanwoo Park, Seungju Han, Xingzhi Guo, Asuman E. Ozdaglar, Kaiqing Zhang, Joo-Kyung Kim

Leveraging multi-agentic frameworks to enhance large language models (LLMs) has demonstrated significant potential recently, with most existing studies focusing on prompting and developing workflows with frozen LLMs. In this paper, we aim to further unleash the power of such multi-agentic frameworks for post-training LLMs for better collaboration. Specifically, we develop a new paradigm of Multi-Agent Post-co-training for collaborative LLMs with Reinforcement Learning (MAPoRL). In MAPoRL, multiple LLMs first generate their own responses and engage in discussions to collaboratively enhance the final response output; the final output is then scored by a verifier, where the scores serve as the reward and is maximized through multi-agent RL. Additionally, MAPoRL also reshapes the reward above with additional incentives to encourage corrective and persuasive outputs in the discussions. A key novelty from most existing LLM post-training paradigms is the advocacy of co-training multiple LLMs together, and the use of RL for better generalization. Accompanied by a few analytical insights, our experiments show that training single LLMs solely is insufficient for encouraging collaboration, while multi-agent co-training can significantly enhance the collaboration performance across multiple datasets, with generalization to unseen domains, compared to that of multiple LLMs before post-training.

Teaching Text Agents to Learn Sequential Decision Making from Failure

Canasai Kruengkrai, Koichiro Yoshino

Text-based reinforcement-learning agents improve their policies by interacting with their environments to collect more training data. However, these self-collected data inevitably contain intermediate failed actions caused by attempting physically infeasible behaviors and/or hallucinations. Directly learning a policy from such trajectories can reinforce incorrect behaviors and reduce task success rates. In this paper, we propose a failed action-aware objective that suppresses the negative impact of failed actions during training by assigning zero return based on textual feedback. Building on this objective, we introduce a perturbation method that leverages unsuccessful trajectories to construct new successful ones that share the same goal. This allows agents to benefit from diverse experiences without further interaction with the environment. Experiments in ALFWorld and ScienceWorld demonstrate that our method significantly outperforms strong baselines and generalizes across environments. Code is available at https://github.com/riken-grp/text-agent.

ScaleBiO: Scalable Bilevel Optimization for LLM Data Reweighting

Rui Pan, Dylan Zhang, Hanning Zhang, Xingyuan Pan, Minrui Xu, Jipeng Zhang, Renjie Pi, Xiaoyu Wang, Tong Zhang

Bilevel optimization has shown its utility across various machine learning settings, yet most algorithms in practice require second-order information, making it challenging to scale them up. Only recently, a paradigm of first-order algorithms has emerged in the theoretical literature, capable of effectively addressing bilevel optimization problems. Nevertheless, the practical efficiency of this paradigm remains unverified, particularly in the context of large language models (LLMs). This paper introduces the first scalable instantiation of this paradigm called _ScaleBiO_, focusing on bilevel optimization for large-scale LLM data reweighting. By combining with a recently proposed memory-efficient training technique called LISA, our novel algorithm allows the paradigm to scale to \sim30B-sized LLMs on 8\times H100 GPUs, marking the first successful application of bilevel optimization under practical scenarios for large-sized LLMs. Empirically, extensive experiments on data reweighting verify the effectiveness of ScaleBiO for different-scaled models, including Llama-3-8B, Gemma-2-9B, Qwen-2-7B, and Qwen-2.5-32B, where bilevel optimization succeeds in instruction-following and math reasoning tasks, outperforming several popular baselines, including uniform sampling, influence-aware data filtering, and reference-model-based sampling methods. Theoretically, ScaleBiO ensures the optimality of the learned data weights, along with a convergence guarantee matching the conventional first-order bilevel optimization paradigm on smooth and strongly convex objectives.

Session 7: Multilinguality and Language Diversity (Tuesday 29 July, 10:30-12:00, Hall 4/5)

Code-Switching Curriculum Learning for Multilingual Transfer in LLMs

Haneul Yoo, Cheonbok Park, Sangdoo Yun, Alice Oh, Hwaran Lee

Large language models (LLMs) now exhibit near human-level performance in various tasks, but their performance drops drastically after a handful of high-resource languages due to the imbalance in pre-training data. Inspired by the human process of second language acquisition, particularly code-switching---the practice of language alternation in a conversation---we propose code-switching curriculum learning (CSCL) to enhance cross-lingual transfer for LLMs. CSCL mimics the stages of human language learning by progressively training models with a curriculum consisting of 1) token-level code-switching, 2) sentence-level code-switching, and 3) monolingual corpora. Using Qwen 2 as our underlying model, we demonstrate the efficacy of the CSCL in improving language transfer to Korean, achieving significant performance gains compared to monolingual continual pre-training methods. Ablation studies reveal that both token- and sentence-level code-switching significantly enhance cross-lingual transfer and that curriculum learning amplifies these effects. We also extend our findings into various languages, including Japanese (high-resource) and Indonesian (low-resource), and using two additional models (Gemma 2 and Phi 3.5). We further show that CSCL mitigates spurious correlations between language resources and safety alignment, presenting a robust, efficient framework for more equitable language transfer in LLMs. We observe that CSCL is effective for low-resource settings where high-quality, monolingual corpora for language transfer are hardly available.

MultiSocial: Multilingual Benchmark of Machine-Generated Text Detection of Social-Media Texts

Dominik Macko, Jakub Kopál, Robert Moro, Ivan Srba

Recent LLMs are able to generate high-quality multilingual texts, indistinguishable for humans from authentic human-written ones. Research in machine-generated text detection is however mostly focused on the English language and longer texts, such as news articles, scientific papers or student essays. Social-media texts are usually much shorter and often feature informal language, grammatical errors, or distinct linguistic items (e.g., emoticons, hashtags). There is a gap in studying the ability of existing methods in detection of such texts, reflected also in the lack of existing multilingual benchmark datasets. To fill this gap we propose the first multilingual (22 languages) and multi-platform (5 social media platforms) dataset for benchmarking machine-generated text detection in the social-media domain, called MultiSocial. It contains 472,097 texts, of which about 58k are human-written and approximately the same amount is generated by each of 7 multilingual LLMs. We use this benchmark to compare existing detection methods in zero-shot as well as fine-tuned form. Our results indicate that the fine-tuned detectors have no problem to be trained on social-media texts and that the platform selection for training matters.

Modular Sentence Encoders: Separating Language Specialization from Cross-Lingual Alignment

Yongxin Huang, Kexin Wang, Goran Glavaš, Iryna Gurevych

Multilingual sentence encoders (MSEs) are commonly obtained by training multilingual language models to map sentences from different languages into a shared semantic space. As such, they are subject to curse of multilinguality, a loss of monolingual representational accuracy due to parameter sharing. Another limitation of MSEs is the trade-off between different task performance: cross-lingual alignment training distorts the optimal monolingual structure of semantic spaces of individual languages, harming the utility of sentence embeddings in monolingual tasks; cross-lingual tasks, such as cross-lingual semantic similarity and zero-shot transfer for sentence classification, may also require conflicting cross-lingual alignment strategies. In this work, we address both issues by means of modular training of sentence encoders. We first train language-specific monolingual modules to mitigate negative interference between languages (i.e., the curse). We then align all non-English sentence embeddings to the English by training cross-lingual alignment adapters, preventing interference with monolingual specialization from the first step. We train the cross-lingual adapters with two different types of data to resolve the conflicting requirements of different cross-lingual tasks. Monolingual and cross-lingual results on semantic text similarity and relatedness, bitext mining and sentence classification show that our modular solution achieves better and more balanced performance across all the tasks compared to full-parameter training of monolithic multilingual sentence encoders, especially benefiting low-resource languages.

Revisiting Common Assumptions about Arabic Dialects in NLP

Amr Keleg, Sharon Goldwater, Walid Magdy

Arabic has diverse dialects, where one dialect can be substantially different from the others. In the NLP literature, some assumptions about these dialects are widely adopted (e.g., ``Arabic dialects can be grouped into distinguishable regional dialects") and are manifested in different computational tasks such as Arabic Dialect Identification (ADI). However, these assumptions are not quantitatively verified. We identify four of these assumptions and examine them by extending and analyzing a multi-label dataset, where the validity of each sentence in 11 different country-level dialects is manually assessed by speakers of these dialects. Our analysis indicates that the four assumptions oversimplify reality, and some of them are not always accurate. This in turn might be hindering further progress in different Arabic NLP tasks.

Cross-Lingual Transfer of Cultural Knowledge: An Asymmetric Phenomenon

Chen Zhang, Zhiyuan Liao, Yansong Feng

Despite substantial research efforts evaluating how well large language models (LLMs) handle global cultural diversity, the mechanisms behind their cultural knowledge acquisition, particularly in multilingual settings, remain unclear. We study this question by investigating how cultural knowledge transfers across languages during the language adaptation of LLMs, a process where an LLM is continually pre-trained to learn another language. We introduce an interpretable framework to study this transfer, ensuring training data transparency and controlling transfer effects. Through a study of four non-Anglophonic cultures, we observe bidirectional cultural transfer between English and other high-resource languages, while low-resource languages primarily transfer knowledge to English with limited reverse flow. To explain this asymmetric phenomenon, we propose a frequency-based hypothesis: cultural knowledge appearing more frequently in the pretraining data transfers more easily, which is supported by empirical analysis of the training corpora. We hope our findings could inform future research on knowledge transfer and promote the development of culturally aware models, particularly for low-resource languages.

Read it in Two Steps: Translating Extremely Low-Resource Languages with Code-Augmented Grammar Books

Chen Zhang, Jiuheng Lin, Xiao Liu, Zekai Zhang, Yansong Feng

While large language models (LLMs) have shown promise in translating extremely low-resource languages using resources like dictionaries, the effectiveness of grammar books remains debated. This paper investigates the role of grammar books in translating extremely low-resource languages by decomposing it into two key steps: grammar rule retrieval and application. To facilitate the study, we introduce ZhuangRules, a modularized dataset of grammar rules and their corresponding test sentences. Our analysis reveals that rule retrieval constitutes a primary bottleneck in grammar-based translation. Moreover, although LLMs can apply simple rules for translation when explicitly provided, they encounter difficulties in handling more complex rules. To address these challenges, we propose representing grammar rules as code functions, considering their similarities in structure and the benefit of code in facilitating LLM reasoning. Our experiments show that using code rules significantly boosts both rule retrieval and application, ultimately resulting in a 13.1% BLEU improvement in translation.

ShifCon: Enhancing Non-Dominant Language Capabilities with a Shift-based Contrastive Framework

Hengyuan Zhang, Chenming Shang, Sizhe Wang, Dongdong Zhang, Feng Yao, Renliang Sun, Yiyao Yu, Yujiu Yang, Furu Wei

Although fine-tuning Large Language Models (LLMs) with multilingual data can rapidly enhance the multilingual capabilities of LLMs, they still exhibit a performance gap between the dominant language (e.g., English) and non-dominant ones due to the imbalance of training data across languages. To further enhance the performance of non-dominant languages, we propose ShifCon, a Shift-based multilingual Contrastive framework that aligns the internal forward process of other languages toward that of the dominant one. Specifically, it shifts the representations of non-dominant languages into the dominant language subspace, allowing them to access relatively rich information encoded in the model parameters. The enriched representations are then shifted back into their original language subspace before generation. Moreover, we introduce a subspace distance metric to pinpoint the optimal layer area for shifting representations and employ multilingual contrastive learning to further enhance the alignment of representations within this area. Experiments demonstrate that our ShifCon framework significantly enhances the performance of non-dominant languages, particularly for low-resource ones. Further analysis offers extra insights to verify the effectiveness of ShifCon and propel future research.

Improving Parallel Sentence Mining for Low-Resource and Endangered Languages

Shu Okabe, Katharina Hämmerl, Alexander Fraser

While parallel sentence mining has been extensively covered for fairly well-resourced languages, pairs involving low-resource languages have received comparatively little attention. To address this gap, we present Belopsem, a benchmark of new datasets for parallel sentence mining on three language pairs where the source side is low-resource and endangered: Occitan-Spanish, Upper Sorbian-German, and Chuvash-Russian. These combinations also reflect varying linguistic similarity within each pair. We compare three language models in an established parallel sentence mining pipeline and apply two types of improvements to one of them, Glot500. We observe better mining quality overall by both applying alignment post-processing with an unsupervised aligner and using a cluster-based isotropy enhancement technique. These findings are crucial for optimising parallel data extraction for low-resource languages in a realistic way.

Can Uniform Meaning Representation Help GPT-4 Translate from Indigenous Languages?

Shira Wein

While ChatGPT and GPT-based models are able to effectively perform many tasks without additional fine-tuning, they struggle with tasks related to extremely low-resource languages and indigenous languages. Uniform Meaning Representation (UMR), a semantic representation designed to capture the meaning of texts in many languages, is well-positioned to be leveraged in the development of low-resource language technologies. In this work, we explore the downstream utility of UMR for low-resource languages by incorporating it into GPT-4 prompts. Specifically, we examine the ability of GPT-4 to perform translation from three indigenous languages (Navajo, Arápaho, and Kukama), with and without demonstrations, as well as with and without UMR annotations. Ultimately, we find that in the majority of our test cases, integrating UMR into the prompt results in a statistically significant increase in performance, which is a promising indication of future applications of the UMR formalism.

x-SAL: Leading Symbolic Reasoning across Languages via Cross-lingual Symbolic-Aided Language Model

Leonardo Ranaldi, Giulia Pucci

Reasoning is an intricate process that transcends both language and vision; yet, despite its inherently modality-agnostic nature, develop- ing effective multilingual and multimodal reasoning capabilities remains a substantial challenge for Multimodal Large Language Models (MLLMs). They struggle to activate complex reasoning behaviours, delivering step-wise explanation, questioning and reflection, particularly in multilingual settings where high-quality supervision across languages is lacking. Recent works have introduced eclectic strategies to enhance MLLMs' reasoning; however, they remain related to a single language. To make MLLMs' reasoning capabilities aligned among languages and improve modality performances, we propose R2-MultiOmnia, a modular approach that instructs the models to abstract key elements of the reasoning process and then refine reasoning trajectories via self-correction. Specifically, we instruct the models producing multimodal synthetic resources by bridging modalities and then self-improving their capabilities. To stabilise learning and the reasoning processes structure, we propose Curriculum Learning Reasoning Stabilisation with structured output rewards to gradually refine the models' capabilities to learn and deliver robust reasoning processes. Experiments show that R2-MultiOmnia improves multimodal reasoning, gets aligned performances among the languages approaching strong models.

Cross-Lingual Pitfalls: Automatic Probing Cross-Lingual Weakness of Multilingual Large Language Models

Zixiang Xu, Yanbo Wang, Yue Huang, Xiuying Chen, Jieyu Zhao, Meng Jiang, Xiangliang Zhang

Large Language Models (LLMs) have achieved remarkable success in Natural Language Processing (NLP), yet their cross-lingual consistency remains a significant challenge. This paper introduces a novel methodology for efficiently identifying inherent cross-lingual weaknesses in LLMs. Our approach leverages beam search and LLM-based simulation to generate bilingual question pairs that expose performance discrepancies between English and target languages. We construct a new dataset of over 6,000 bilingual pairs across 16 languages using this methodology, demonstrating its effectiveness in revealing weaknesses even in state-of-the-art models. The extensive experiments demonstrate that our method precisely and cost-effectively pinpoints cross-lingual weaknesses, consistently revealing over 50% accuracy drops in target languages across a wide range of models. Moreover, further experiments investigate the relationship between linguistic similarity and cross-lingual weaknesses, revealing that linguistically related languages share similar performance patterns and benefit from targeted post-training. Code is available at https://github.com/xzx34/Cross-Lingual-Pitfalls.

PPT: A Minor Language News Recommendation Model via Cross-Lingual Preference Pattern Transfer

Yiyang Zhang, Nan Chen

Rich user-item interactions are essential for building reliable recommender systems, as they reflect user preference patterns. However, minor language news recommendation platforms suffer from limited interactions due to a small user base. A natural solution is to apply well-established English recommender systems to minor language news recommendation, but the linguistic gap can lead to inaccurate modeling of minor language news content. Therefore, enabling few-shot minor language news recommender systems to capture both content information and preference patterns remains a challenge. Based on the observation that preference patterns are similar across languages, we propose a minor language news recommendation model by cross-lingual preference pattern transfer, named PPT. Our model adopts the widely used two-tower architecture and employs the large language model as the backbone of the news encoder. Through cross-lingual alignment, the strong English capability of the news encoder is extended to minor languages, thus enhancing news content representations. Additionally, through cross-lingual news augmentation, PPT simulates interactions of minor language news in the English domain, which facilitates the transfer of preference patterns from the many-shot English domain to the few-shot minor language domain. Extensive experiments on two real-world datasets across 15 minor languages demonstrate the superiority and generalization of our proposed PPT in addressing minor language news recommendation.

Middle-Layer Representation Alignment for Cross-Lingual Transfer in Fine-Tuned LLMs

Danni Liu, Jan Niehues

While large language models demonstrate remarkable capabilities at task-specific applications through fine-tuning, extending these benefits across diverse languages is essential for broad accessibility. However, effective cross-lingual transfer is hindered by LLM performance gaps across languages and the scarcity of fine-tuning data in many languages. Through analysis of LLM internal representations from over 1,000+ language pairs, we discover that middle layers exhibit the strongest potential for cross-lingual alignment. Building on this finding, we propose a middle-layer alignment objective integrated into task-specific training. Our experiments on slot filling, machine translation, and structured text generation show consistent improvements in cross-lingual transfer, especially to lower-resource languages. The method is robust to the choice of alignment languages and generalizes to languages unseen during alignment. Furthermore, we show that separately trained alignment modules can be merged with existing task-specific modules, improving cross-lingual capabilities without full re-training. The code is provided in the supplementary materials.

Less, but Better: Efficient Multilingual Expansion for LLMs via Layer-wise Mixture-of-Experts

Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Yufeng Chen, Jinan Xu, Jie Zhou

Continually expanding new languages for existing large language models (LLMs) is a promising yet challenging approach to building powerful multilingual LLMs. The biggest challenge is to make the model continuously learn new languages while preserving the proficient ability of old languages. To achieve this, recent work utilizes the Mixture-of-Experts (MoE) architecture to expand new languages by adding new experts and avoid catastrophic forgetting of old languages by routing corresponding tokens to the original model backbone (old experts). Although intuitive, this kind of method is parameter-costly when expanding new languages and still inevitably impacts the performance of old languages. To address these limitations, we analyze the language characteristics of different layers in LLMs and propose a layer-wise expert allocation algorithm (LayerMoE) to determine the appropriate number of new experts for each layer. Specifically, we find different layers in LLMs exhibit different representation similarities between languages and then utilize the similarity as the indicator to allocate experts for each layer, i.e., the higher similarity, the fewer experts. Additionally, to further mitigate the forgetting of old languages, we add a classifier in front of the router network on the layers with higher similarity to guide the routing of old language tokens. Experimental results show that our method outperforms the previous state-of-the-art baseline with 60\% fewer experts in the single-expansion setting and with 33.3\% fewer experts in the lifelong-expansion setting, demonstrating the effectiveness of our method.

Minimal Pair-Based Evaluation of Code-Switching

Igor Sterner, Simone Teufel

There is a lack of an evaluation methodology that estimates the extent to which large language models (LLMs) use code-switching (CS) in the same way as bilinguals. Existing methods do not have wide language coverage, fail to account for the diverse range of CS phenomena, or do not scale. We propose an intervention based on minimal pairs of CS. Each minimal pair contains one naturally occurring CS sentence and one minimally manipulated variant. We collect up to 1,000 such pairs each for 11 language pairs. Our human experiments show that, for every language pair, bilinguals consistently prefer the naturally occurring CS sentence. Meanwhile our experiments with current LLMs show that the larger the model, the more consistently it assigns higher probability to the naturally occurring CS sentence than to the variant. In accordance with theoretical claims, the largest probability differences arise in those pairs where the manipulated material consisted of closed-class words.

Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, Jian Gang Ngui, David Ifeoluwa Adelani, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Daphne Ippolito, Enzo Ferrante, Leshem Choshen, Marzieh Fadaee, Beyza Ermis, Sara Hooker

Reliable multilingual evaluation is difficult, and culturally appropriate evaluation is even harder to achieve. A common practice to fill this gap is to machine-translate English evaluation sets. However, translation introduces language bias and carries over cultural and regional assumptions from the original questions -- often testing knowledge irrelevant to the target audience. In this work, we highlight the extent and impact of these biases and present a multilingual evaluation framework that aims to mitigate them through improved translations and annotation practices. Through a large-scale study involving professional and community translators and annotators, we show that state-of-the-art models excel primarily by learning Western-centric concepts. Notably, we find that model rankings on the full MMLU change when evaluated on a subset of questions explicitly marked as culturally sensitive. We release Global MMLU, a multilingual extension of MMLU across 42 languages, featuring improved translation quality, expanded language coverage, and designated subsets labeled as culturally sensitive and culturally agnostic to enable a more comprehensive and equitable benchmark for evaluating language models across diverse linguistic and cultural contexts.

CC-Tuning: A Cross-Lingual Connection Mechanism for Improving Joint Multilingual Supervised Fine-Tuning

Yangfan Ye, Xiaocheng Feng, Zekun Yuan, Xiachong Feng, Libo Qin, Lei Huang, Weitao Ma, Yichong Huang, Zhirui Zhang, Yunfei Lu, Xiaohui Yan, Duyu Tang, Dandan Tu, Bing Qin

Current large language models (LLMs) often exhibit imbalanced multilingual capabilities due to their English-centric training corpora. To address this, existing fine-tuning approaches operating at the data-level (e.g., through data augmentation or distillation) typically introduce implicit cross-lingual alignment, overlooking the potential for more profound, latent-level cross-lingual interactions. In this work, we propose CC-Tuning, a novel multilingual fine-tuning paradigm that explicitly establishes a cross-lingual connection mechanism at the latent level. During training, CC-Tuning fuses the feed forward activations from both English and non-English inputs, enabling the model to benefit from both linguistic resources. This process is facilitated with a trainable Decision Maker that identifies beneficial activations. Furthermore, during inference, a Transform Matrix is utilized to simulate the cross-lingual connection under monolingual setting through representation transformation. Our experiments on six benchmarks covering 22 languages show that CC-Tuning outperforms vanilla SFT and offers a strong latent-level alternative to data-level augmentation methods. Further analysis also highlights the practicality of CC-Tuning and the potential of latent-level cross-lingual interactions in advancing the multilingual performance of LLMs.

Multilingual Arbitration: Optimizing Data Pools to Accelerate Multilingual Progress

Ayomide Odumakinde, Daniel D'souza, Pat Verga, Beyza Ermis, Sara Hooker

Synthetic data has driven recent state-of-the-art advancements, but reliance on a single oracle teacher model can lead to model collapse and bias propagation. These issues are particularly severe in multilingual settings, where no single model excels across all languages. In this study, we propose multilingual arbitration, which exploits performance variations among multiple models for each language. By strategically routing samples through a diverse set of models, each with unique strengths, we mitigate these challenges and enhance multilingual performance. Extensive experiments with state-of-the-art models demonstrate that our approach significantly surpasses single-teacher distillation, achieving up to 80% win rates over proprietary and open-weight models like Gemma 2, Llama 3.1, and Mistral v0.3, with the largest improvements in low-resource languages.

That doesn't sound right: Evaluating speech transcription quality in field linguistics corpora

Eric Le Ferrand, Bo Jiang, Emily Prud'hommeaux, Joshua Hartshorne

Incorporating automatic speech recognition (ASR) into field linguistics workflows for language documentation has become increasingly common. While ASR performance has seen improvements in low-resource settings, obstacles remain when training models on data collected by documentary linguists. One notable challenge lies in the way that this data is curated. ASR datasets built from spontaneous speech are typically recorded in consistent settings and transcribed by native speakers following a set of well designed guidelines. In contrast, field linguists collect data in whatever format it is delivered by their language consultants and transcribe it as best they can given their language skills and the quality of the recording. This approach to data curation, while valuable for linguistic research, does not always align with the standards required for training robust ASR models. In this paper, we explore methods for identifying speech transcriptions in fieldwork data that may be unsuitable for training ASR models. We focus on two complimentary automated measures of transcription quality that can be used to identify transcripts with characteristics that are common in field data but could be detrimental to ASR training. We show that one of the metrics is highly effective at retrieving these types of transcriptions. Additionally, we find that filtering datasets using this metric of transcription quality reduces WER both in controlled experiments using simulated fieldwork with artificially corrupted data and in real fieldwork corpora.

Disentangling Language Medium and Culture Context for Evaluating Multilingual Large Language Models

Jiahao Ying, Wei Tang, Yiran Zhao, Yixin Cao, Yu Rong, Wenxuan Zhang

This paper introduces a Dual Evaluation Framework to comprehensively assess the multilingual capabilities of LLMs. By decomposing the evaluation along the dimensions of linguistic medium and cultural context, this framework enables a nuanced analysis of LLMs' ability to process questions within both native and cross-cultural contexts cross-lingually. Extensive evaluations are conducted on a wide range of models, revealing a notable "Cultural-Linguistic Synergy" phenomenon, where models exhibit better performance when questions are culturally aligned with the language. This phenomenon is further explored through interpretability probing, which shows that a higher proportion of specific neurons are activated in a language's cultural context. This activation proportion could serve as a potential indicator for evaluating multilingual performance during model training. Our findings challenge the prevailing notion that LLMs, primarily trained on English data, perform uniformly across languages and highlight the necessity of culturally and linguistically model evaluations.

TUMLU: A Unified and Native Language Understanding Benchmark for Turkic Languages

Jafar Isbarov, Arofat Akhundjanova, Mammad Hajili, Kavsar Huseynova, Dmitry Gaynullin, Anar Rzayev, Osman Tursun, Aizirek Turdubaeva, Ilshat Saetov, Rinat Kharisov, Saule Belginova, Ariana Kenbayeva, Amina Alisheva, Abdullatif Köksal, SAMIR RUSTAMOV, Duygu Ataman

Being able to thoroughly assess massive multi-task language understanding (MMLU) capabilities is essential for advancing the applicability of multilingual language models. However, preparing such benchmarks in high quality native language is often costly and therefore limits the representativeness of evaluation datasets. While recent efforts focused on building more inclusive MMLU benchmarks, these are conventionally built using machine translation from high-resource languages, which may introduce errors and fail to account for the linguistic and cultural intricacies of the target languages. In this paper, we address the lack of native language MMLU benchmark especially in the under-represented Turkic language family with distinct morphosyntactic and cultural characteristics. We propose two benchmarks for Turkic language MMLU: TUMLU is a comprehensive, multilingual, and natively developed language understanding benchmark specifically designed for Turkic languages. It consists of middle- and high-school level questions spanning 11 academic subjects in Azerbaijani, Crimean Tatar, Karakalpak, Kazakh, Kyrgyz, Tatar, Turkish, Uyghur, and Uzbek. We also present TUMLU-mini, a more concise, balanced, and manually verified subset of the dataset. Using this dataset, we systematically evaluate a diverse range of open and proprietary multilingual large language models (LLMs), including Claude, Gemini, GPT, and LLaMA, offering an in-depth analysis of their performance across different languages, subjects, and alphabets. To promote further research and development in multilingual language understanding, we release TUMLU-mini and all corresponding evaluation scripts.

From English to Second Language Mastery: Enhancing LLMs with Cross-Lingual Continued Instruction Tuning

Linjuan Wu, Hao-Ran Wei, Baosong Yang, Weiming Lu

Supervised Fine-Tuning (SFT) with translated instruction data effectively adapts Large Language Models (LLMs) from English to non-English languages. We introduce Cross-Lingual Continued Instruction Tuning (X-CIT), which fully leverages translation-based parallel instruction data to enhance cross-lingual adaptability. X-CIT emulates the human process of second language acquisition and is guided by Chomsky's Principles and Parameters Theory. It first fine-tunes the LLM on English instruction data to establish foundational capabilities (i.e. Principles), then continues with target language translation and customized chat-instruction data to adjust "parameters" specific to the target language. This chat-instruction data captures alignment information in translated parallel data, guiding the model to initially think and respond in its native language before transitioning to the target language. To further mimic human learning progression, we incorporate Self-Paced Learning (SPL) during continued training, allowing the model to advance from simple to complex tasks. Implemented on Llama-2-7B across five languages, X-CIT was evaluated against three objective benchmarks and an LLM-as-a-judge benchmark, improving the strongest baseline by an average of 1.97% and 8.2% in these two benchmarks, respectively.

Analyzing LLMs' Cognition of Knowledge Boundary Across Languages Through the Lens of Internal Representation

Chenghao Xiao, Hou Pong Chan, Hao Zhang, Mahani Aljunied, Lidong Bing, Noura Al Moubayed, Yu Rong

While understanding the knowledge boundaries of LLMs is crucial to prevent hallucination, research on the knowledge boundaries of LLMs has predominantly focused on English. In this work, we present the first study to analyze how LLMs recognize knowledge boundaries across different languages by probing their internal representations when processing known and unknown questions in multiple languages. Our empirical studies reveal three key findings: 1) LLMs' perceptions of knowledge boundaries are encoded in the middle to middle-upper layers across different languages. 2) Language differences in knowledge boundary perception follow a linear structure, which motivates our proposal of a training-free alignment method that effectively transfers knowledge boundary perception ability across languages, thereby helping reduce hallucination risk in low-resource languages; 3) Fine-tuning on bilingual question pair translation further enhances LLMs' recognition of knowledge boundaries across languages. Given the absence of standard testbeds for cross-lingual knowledge boundary analysis, we construct a multilingual evaluation suite comprising three representative types of knowledge boundary data. Our code and datasets are publicly available at .

Language Fusion for Parameter-Efficient Cross-lingual Transfer

Philipp Borchert, Ivan Vulić, Marie-Francine Moens, Jochen De Weerdt

Limited availability of multilingual text corpora for training language models often leads to poor performance on downstream tasks due to undertrained representation spaces for languages other than English. This 'under-representation' has motivated recent cross-lingual transfer methods to leverage the English representation space by e.g. mixing English and 'non-English' tokens at the input level or extending model parameters to accommodate new languages. However, these approaches often come at the cost of increased computational complexity. We propose Fusion for Language Representations (FLARE) in adapters, a novel method that enhances representation quality and downstream performance for languages other than English while maintaining parameter efficiency. FLARE integrates source and target language representations within low-rank (LoRA) adapters using lightweight linear transformations, maintaining parameter efficiency while improving transfer performance. A series of experiments across representative cross-lingual natural language understanding tasks, including natural language inference, question-answering and sentiment analysis, demonstrate FLARE's effectiveness. FLARE achieves performance improvements of 4.9% for Llama 3.1 and 2.2% for Gemma 2 compared to standard LoRA fine-tuning on question-answering tasks, as measured by the exact match metric.

Do Language Models Understand Honorific Systems in Javanese?

Mohammad Rifqi Farhansyah, Iwan Darmawan, Adryan Kusumawardhana, Genta Indra Winata, Alham Fikri Aji, Derry Tanti Wijaya

The Javanese language features a complex system of honorifics that vary according to the social status of the speaker, listener, and referent. Despite its cultural and linguistic significance, there has been limited progress in developing a comprehensive corpus to capture these variations for natural language processing (NLP) tasks. In this paper, we present Unggah-Ungguh, a carefully curated dataset designed to encapsulate the nuances of Unggah-Ungguh Basa, the Javanese speech etiquette framework that dictates the choice of words and phrases based on social hierarchy and context. Using Unggah-Ungguh, we assess the ability of language models (LMs) to process various levels of Javanese honorifics through classification and machine translation tasks. To further evaluate cross-lingual LMs, we conduct machine translation experiments between Javanese (at specific honorific levels) and Indonesian. Additionally, we explore whether LMs can generate contextually appropriate Javanese honorifics in conversation tasks, where the honorific usage should align with the social role and contextual cues. Our findings indicate that current LMs struggle with most honorific levels, exhibiting a bias toward certain honorific tiers.

NusaAksara: A Multimodal and Multilingual Benchmark for Preserving Indonesian Indigenous Scripts

Muhammad Farid Adilazuarda, Musa Izzanardi Wijanarko, Lucky Susanto, Khumaisa Nur'aini, Derry Tanti Wijaya, Alham Fikri Aji

Indonesia is rich in languages and scripts. However, most NLP progress has been made using romanized text. In this paper, we present NusaAksara, a novel public benchmark for Indonesian languages that includes their original scripts. Our benchmark covers both text and image modalities and encompasses diverse tasks such as image segmentation, OCR, transliteration, translation, and language identification. Our data is constructed by human experts through rigorous steps. NusaAksara covers 8 scripts across 7 languages, including low-resource languages not commonly seen in NLP benchmarks. Although unsupported by Unicode, the Lampung script is included in this dataset. We benchmark our data across several models, from LLMs and VLMs such as GPT-4o, Llama 3.2, and Aya 23 to task-specific systems such as PP-OCR and LangID, and show that most NLP technologies cannot handle Indonesia's local scripts, with many achieving near-zero performance.

Truth Knows No Language: Evaluating Truthfulness Beyond English

Blanca Calvo Figueras, Eneko Sagarzazu, Julen Etxaniz, Jeremy Barnes, Pablo Gamallo, Iria de-Dios-Flores, Rodrigo Agerri

We introduce a professionally translated extension of the TruthfulQA benchmark designed to evaluate truthfulness in Basque, Catalan, Galician, and Spanish. Truthfulness evaluations of large language models (LLMs) have primarily been focused on English. However, the ability of LLMs to maintain truthfulness across languages remains under-explored. Our study evaluates 12 state-of-the-art open LLMs, comparing base and instruction-tuned models using human evaluation, multiple-choice metrics, and LLM-as-a-Judge scoring. Our findings reveal that, while LLMs perform best in English and worst in Basque (the lowest-resourced language), overall truthfulness discrepancies across languages are smaller than anticipated. Furthermore, we show that LLM-as-a-Judge correlates more closely with human judgments than multiple-choice metrics, and that informativeness plays a critical role in truthfulness assessment. Our results also indicate that machine translation provides a viable approach for extending truthfulness benchmarks to additional languages, offering a scalable alternative to professional translation. Finally, we observe that universal knowledge questions are better handled across languages than context- and time-dependent ones, highlighting the need for truthfulness evaluations that account for cultural and temporal variability. Datasets, models and code are publicly available under open licenses.

Where Are We? Evaluating LLM Performance on African Languages

Ife Adebara, Hawau Olamide Toyin, Nahom Tesfu Ghebremichael, AbdelRahim A. Elmadany, Muhammad Abdul-Mageed

Africa's rich linguistic heritage remains underrepresented in NLP, largely due to historical policies that favor foreign languages and create significant data inequities. In this paper, we integrate theoretical insights on Africa's language landscape with an empirical evaluation using Sahara— a comprehensive benchmark curated from large-scale, publicly accessible datasets capturing the continent's linguistic diversity. By systematically assessing the performance of leading large language models (LLMs) on Sahara, we demonstrate how policy-induced data variations directly impact model effectiveness across African languages. Our findings reveal that while a few languages perform reasonably well, many Indigenous languages remain marginalized due to sparse data. Leveraging these insights, we offer actionable recommendations for policy reforms and inclusive data practices. Overall, our work underscores the urgent need for a dual approach—combining theoretical understanding with empirical evaluation—to foster linguistic diversity in AI for African communities.

AfroCS-xs: Creating a Compact, High-Quality, Human-Validated Code-Switched Dataset for African Languages

Kayode Olaleye, Arturo Oncevay, Mathieu Sibue, Nombuyiselo Zondi, Michelle Terblanche, Sibongile Mapikitla, Richard Lastrucci, Charese Smiley, Vukosi Marivate

Code-switching is prevalent in multilingual communities but lacks adequate high-quality data for model development, especially for African languages. To address this, we present AfroCS-xs, a small human-validated synthetic code-switched dataset for four African languages (Afrikaans, Sesotho, Yoruba, isiZulu) and English within a specific domain—agriculture. Using large language models (LLMs), we generate code-switched sentences, including English translations, that are rigorously validated and corrected by native speakers. As a downstream evaluation task, we use this dataset to fine-tune different instruction-tuned LLMs for code-switched translation and compare their performance against machine translation (MT) models. Our results demonstrate that LLMs consistently improve in translation accuracy when fine-tuned on the high-quality AfroCS-xs dataset, highlighting that substantial gains can still be made with a low volume of data. We also observe improvements on natural code-switched and out-of-domain (personal finance) test sets. Overall, regardless of data size and prior exposure to a language, LLMs benefit from higher quality training data when translating code-switched texts in under-represented languages.

Session 7: Multimodality and Language Grounding to Vision, Robotics and Beyond (Tuesday 29 July, 10:30-12:00, Hall 4/5)

LlamaV-o1: Rethinking Step-by-step Visual Reasoning in LLMs

Omkar Thawakar, Dinura Dissanayake, Ketan Pravin More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Ilmuz Zaman Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, Hisham Cholakkal, Ivan Laptev, Mubarak Shah, Fahad Shahbaz Khan, Salman Khan

Step-by-step reasoning is crucial for solving complex visual tasks, yet existing approaches lack a comprehensive framework for evaluating this capability and do not emphasize step-wise problem-solving. To this end, we propose a comprehensive framework for advancing multi-step visual reasoning in large multimodal models (LMMs) through three key contributions. First, we introduce a Visual Reasoning Chain Benchmark, the most comprehensive benchmark for multi-step visual reasoning, covering eight diverse categories and over 4k reasoning steps. This enables rigorous evaluation of LMMs' ability to reason accurately and interpretably across multiple steps. Second, we propose a fine-grained reasoning metric that evaluates correctness and logical coherence at each step, providing deeper insights beyond traditional accuracy metrics. Third, we introduce LlamaV-o1, a state-of-the-art multimodal reasoning model trained using a multi-step curriculum learning approach. LlamaV-o1 is optimized for structured, step-by-step reasoning and significantly outperforms existing open-source models. It surpasses Llava-CoT with a 3.8% absolute gain across six benchmarks, achieving an average score of 67.3 while being 5x faster during inference scaling. Our benchmark, model, and code is available at https://github.com/mbzuai-oryx/LlamaV-o1.

Session 7: NLP Applications (Tuesday 29 July, 10:30-12:00, Hall 4/5)

Structure-adaptive Adversarial Contrastive Learning for Multi-Domain Fake News Detection

Lingwei Wei, Dou Hu, Wei Zhou, Philip S. Yu, Songlin Hu

The rapid proliferation of fake news across multiple domains poses significant threats to society. Existing multi-domain detection models typically capture domain-shared semantic features to achieve generalized detection. However, they often fail to generalize well due to poor adaptability, which limits their ability to provide complementary features for detection, especially in data-constrained conditions. To address these challenges, we investigate the propagation-adaptive multi-domain fake news detection paradigm. We propose a novel framework, Structure-adaptive Adversarial Contrastive Learning (StruACL), to adaptively enable structure knowledge transfer between multiple domains. Specifically, we first contrast representations between content-only and propagation-rich data to preserve structural patterns in the shared representation space. Additionally, we design a propagation-guided adversarial training strategy to enhance the diversity of representations. Under the StruACL objective, we leverage a unified Transformer-based and graph-based model to jointly learn transferable semantic and structural features for detection across multiple domains. Experiments on seven fake news datasets demonstrate that StruACL-TGN achieves better multi-domain detection performance on general and data-constrained scenarios, showing the effectiveness and better generalization of StruACL.

Explainable Depression Detection in Clinical Interviews with Personalized Retrieval-Augmented Generation

Linhai Zhang, Ziyang Gao, Deyu Zhou, Yulan He

Depression is a widespread mental health disorder, and clinical interviews are the gold standard for assessment. However, their reliance on scarce professionals highlights the need for automated detection. Current systems mainly employ black-box neural networks, which lack interpretability, which is crucial in mental health contexts. Some attempts to improve interpretability use post-hoc LLM generation but suffer from hallucination. To address these limitations, we propose RED, a Retrieval-augmented generation framework for Explainable depression Detection. RED retrieves evidence from clinical interview transcripts, providing explanations for predictions. Traditional query-based retrieval systems use a one-size-fits-all approach, which may not be optimal for depression detection, as user backgrounds and situations vary. We introduce a personalized query generation module that combines standard queries with user-specific background inferred by LLMs, tailoring retrieval to individual contexts. Additionally, to enhance LLM performance in social intelligence, we augment LLMs by retrieving relevant knowledge from a social intelligence datastore using an event-centric retriever. Experimental results on the real-world benchmark demonstrate RED's effectiveness compared to neural networks and LLM-based baselines.

Reasoning Does Not Necessarily Improve Role-Playing Ability

Xiachong Feng, Longxu Dou, Lingpeng Kong

The application of role-playing large language models (LLMs) is rapidly expanding in both academic and commercial domains, driving an increasing demand for high-precision role-playing models. Simultaneously, the rapid advancement of reasoning techniques has continuously pushed the performance boundaries of LLMs. This intersection of practical role-playing demands and evolving reasoning capabilities raises an important research question: Can reasoning techniques enhance the role-playing capabilities of LLMs?" To address this, we conduct a comprehensive study using 6 role-playing benchmarks, 24 LLMs, and 3 distinct role-playing strategies, comparing the effectiveness of direct zero-shot role-playing, role-playing with Chain-of-Thought (CoT), and role-playing using reasoning-optimized LLMs. Our findings reveal that CoT may reduce role-playing performance, reasoning-optimized LLMs are unsuitable for role-playing, reasoning ability disrupts the role-playing scaling law, and large models still lack proficiency in advanced role-playing. Furthermore, based on extensive experimental results, we propose two promising future research directions: Role-aware Chain-of-Thought for improving role-playing LLMs and Reinforcement Learning for role-playing LLMs, aiming to enhance the adaptability, consistency, and effectiveness of role-playing LLMs for both research and real-world applications.

TableLLM: Enabling Tabular Data Manipulation by LLMs in Real Office Usage Scenarios

Xiaokang Zhang, Sijia Luo, Bohan Zhang, Zeyao Ma, Jing Zhang, Yang Li, Guanlin Li, Zijun Yao, Kangli Xu, Jinchang Zhou, Daniel Zhang-Li, Jifan Yu, Shu Zhao, Juanzi Li, Jie Tang

We introduce TableLLM, a robust large language model (LLM) with 8 billion parameters, purpose-built for proficiently handling tabular data manipulation tasks, whether they are embedded within documents or spreadsheets, catering to real-world office scenarios. We propose a distant supervision method for training, which comprises a reasoning process extension strategy, aiding in training LLMs to understand reasoning patterns more effectively as well as a cross-way validation strategy, ensuring the quality of the automatically generated data. To evaluate the performance of TableLLM, we have crafted benchmarks tailored to address both document and spreadsheet formats as well as constructed a well-organized evaluation pipeline capable of handling both scenarios. Thorough evaluations underscore the advantages of TableLLM when compared to various existing general-purpose and tabular data-focused LLMs. We have publicly released the model checkpoint, source code, benchmarks, and a web application for user interaction on this anonymized repository.

A Survey of LLM-based Agents in Medicine

Wenxuan Wang, Zizhan Ma, Zheng WANG, Chenghan Wu, Jiaming Ji, Wenting Chen, Xiang Li, Yixuan Yuan

Large Language Models (LLMs) are transforming healthcare through LLM-based agents that can understand and assist with medical tasks. This survey examines the architectures, applications, and challenges of LLM-based agents in medicine. We analyze key components including system profiles, clinical planning, medical reasoning frameworks, and external capacity enhancement. The survey covers major applications in clinical decision support, medical documentation, training simulations, and healthcare service optimization, along with evaluation frameworks and metrics. While these agents show promise in enhancing healthcare delivery, challenges remain in hallucination management, multimodal integration, implementation, and ethics. We conclude by highlighting future directions in medical reasoning, physical system integration, and training simulations, providing researchers and practitioners with a structured overview of the field's current state and prospects.

Towards Conditioning Clinical Text Generation for User Control

Osman Alperen Koraş, Rabi Bahnan, Jens Kleesiek, Amin Dada

Deploying natural language generation systems in clinical settings remains challenging despite advances in Large Language Models (LLMs), which continue to exhibit hallucinations and factual inconsistencies, necessitating human oversight. This paper explores automated dataset augmentation using LLMs as human proxies to condition LLMs for clinician control without increasing cognitive workload. On the BioNLP ACL'24 Discharge Me! Shared Task, we achieve new state-of-the-art results with simpler methods than prior submissions through more efficient training, yielding a 9% relative improvement without augmented training and up to 34% with dataset augmentation. Preliminary human evaluation further supports the effectiveness of our approach, highlighting the potential of augmenting clinical text generation for control to enhance relevance, accuracy, and factual consistency.

Bridge-Coder: Transferring Model Capabilities from High-Resource to Low-Resource Programming Language

Jipeng Zhang, Jianshu Zhang, Yuanzhe LI, Renjie Pi, Rui Pan, Runtao Liu, Zheng Ziqiang, Tong Zhang

Most LLMs universally excel at generating code for high-resource programming languages (HRPLs) like Python, a capability that has become standard due to the abundance of training data. However, they struggle significantly with low-resource programming languages (LRPLs) such as D, exacerbating the digital divide. This gap limits developers using LRPLs from equally benefiting and hinders innovation within underrepresented programming communities. To make matters worse, manually generating data for LRPLs is highly labor intensive and requires expensive expert effort. In this work, we begin by analyzing the NL-PL Gap, where LLMs' direct-generated LRPL data often suffers from subpar quality due to the misalignment between natural language (NL) instructions and programming language (PL) outputs. To address this issue, we introduce Bridge-Assist Generation, a method to generate LRPL data utilizing LLM's general knowledge, HRPL proficiency, and in-context learning capabilities. To further maximize the utility of the generated data, we propose Bridged Alignment to obtain Bridge-Coder. To thoroughly evaluate our approach, we select four relatively LRPLs: R, D, Racket, and Bash. Experimental results reveal that Bridge-Coder achieves significant improvements over the original model, with average gains of 18.71 and 10.81 on two comprehensive benchmarks, M-HumanEval and M-MBPP.

Self-Improvement Towards Pareto Optimality: Mitigating Preference Conflicts in Multi-Objective Alignment

Moxin Li, Yuantao Zhang, Wenjie Wang, Wentao Shi, Zhuo Liu, Fuli Feng, Tat-Seng Chua

Multi-Objective Alignment (MOA) aims to align LLMs' responses with multiple human preference objectives, with Direct Preference Optimization (DPO) emerging as a prominent approach. However, we find that DPO-based MOA approaches suffer from widespread preference conflicts in the data, where different objectives favor different responses. This results in conflicting optimization directions, hindering the optimization on the Pareto Front. To address this, we propose to construct Pareto-optimal responses to resolve preference conflicts. To efficiently obtain and utilize such responses, we propose a self-improving DPO framework that enables LLMs to self-generate and select Pareto-optimal responses for self-supervised preference alignment. Extensive experiments on two datasets demonstrate the superior Pareto Front achieved by our framework compared to various baselines\footnote{Code is available at .

Beyond Browsing: API-Based Web Agents

Yueqi Song, Frank F. Xu, Shuyan Zhou, Graham Neubig

Web browsers are a portal to the internet, where much of human activity is undertaken. Thus, there has been significant research work in AI agents that interact with the internet through web browsing. However, there is also another interface designed specifically for machine interaction with online content: application programming interfaces (APIs). In this paper we ask - *what if we were to take tasks traditionally tackled by Browsing Agents, and give AI agents access to APIs*? To do so, we propose two varieties of agents: (1) an API-calling agent that attempts to perform online tasks through APIs only, similar to traditional coding agents, and (2) a Hybrid Agent that can interact with online data through both web browsing and APIs. In experiments on WebArena, a widely-used and realistic benchmark for web navigation tasks, we find that API-Based Agents outperform web Browsing Agents. Hybrid Agents out-perform both others nearly uniformly across tasks, resulting in a more than 20.0% absolute improvement over web browsing alone, achieving a success rate of 35.8%, the SOTA performance among task-agnostic agents. These results strongly suggest that when APIs are available, they present an attractive alternative to relying on web browsing alone.

Math2Visual: A Framework for Generating Pedagogically Meaningful Visuals for Teaching Math Word Problems

Junling Wang, Anna Rutkiewicz, April Wang, Mrinmaya Sachan

Visuals are valuable tools in teaching math word problems (MWPs), helping young learners interpret textual descriptions into mathematical expressions before solving them. However, creating such visuals is labor-intensive and we lack automated methods. In this paper, we present Math2Visual, an automatic framework for generating pedagogically meaningful visuals from MWP text descriptions. Math2Visual leverages a pre-defined visual language and a structured design space for visuals, informed by math teachers, to effectively capture the essential mathematical relationships within MWPs. Using Math2Visual, we construct an annotated dataset of 1,903 visuals and evaluate Text-to-Image (TTI) models in generating visuals that align with our design. We further fine-tune various TTI models with our dataset, demonstrating improvements in educational visual generation. Our work establishes a new benchmark for automated pedagogical meaningful visual generation and offers insights into the challenge of generating multimodal educational content.

LEMMA: Learning from Errors for MatheMatical Advancement in LLMs

Zhuoshi Pan, Yu Li, Honglin Lin, Qizhi Pei, Zinan Tang, Wei Wu, Chenlin Ming, H. Vicky Zhao, Conghui He, Lijun Wu

Large language models (LLMs) have demonstrated remarkable reasoning capability in solving mathematical problems. However, existing approaches primarily focus on improving the quality of correct training data, e.g., distilling high-quality correct solutions from advanced models, neglecting the value contained in error data, potentially hindering the model's reflective ability. Though some studies attempted to leverage error data, they often involve complex mechanisms, such as Monte Carlo Tree Search (MCTS) to explore error nodes. In this work, we propose to enhance LLM's reasoning ability by Learning from Errors for MatheMatical Advancement (LEMMA). LEMMA constructs data consists of an incorrect solution with an erroneous step and a reflection connection to a correct solution for fine-tuning. Specifically, we systematically analyze the model-generated error types and introduce an _error-type grounded mistake augmentation_ method to collect diverse and representative errors. Correct solutions are either from fixing the errors or generating a fresh start. By fine-tuning on the constructed dataset, the model is able to _self-correct errors autonomously_ within the generation process _without relying on external critique models_. Experimental results demonstrate that LEMMA achieves significant performance improvements over other strong models with less than 90k data.

P3: Prompts Promote Prompting

Xinyu Zhang, Yuanquan Hu, Fangchao Liu, Zhicheng Dou

Current large language model (LLM) applications often employ multi-component prompts, comprising both system and user prompts, to guide model behaviors. While recent advancements have demonstrated the efficacy of automatically optimizing either the system or user prompt to boost performance, such unilateral approaches often yield suboptimal outcomes due to the interdependent nature of these components. In this work, we introduce P3, a novel self-improvement framework that concurrently optimizes both system and user prompts through an iterative process. The offline optimized prompts are further leveraged to promote online prompting by performing query-dependent prompt optimization. Extensive experiments on general tasks (e.g., Arena-hard and Alpaca-eval) and reasoning tasks (e.g., GSM8K and GPQA) demonstrate that P3 achieves superior performance in the realm of automatic prompt optimization. Our results highlight the effectiveness of a holistic optimization strategy in enhancing LLM performance across diverse domains.

Quantile Regression with Large Language Models for Price Prediction

Nikhita Vedula, Dushyanta Dhyani, Laleh Jalali, Boris N. Oreshkin, Mohsen Bayati, Shervin Malmasi

Large Language Models (LLMs) have shown promise in structured prediction tasks, including regression, but existing approaches primarily focus on point estimates and lack systematic comparison across different methods. We investigate probabilistic regression using LLMs for unstructured inputs, addressing challenging text-to-distribution prediction tasks such as price estimation where both nuanced text understanding and uncertainty quantification are critical. We propose a novel quantile regression approach that enables LLMs to produce full predictive distributions, improving upon traditional point estimates. Through extensive experiments across three diverse price prediction datasets, we demonstrate that fine-tuned Mistral-7B with quantile heads significantly outperforms traditional approaches for both point and distributional estimations, as measured by three established metrics each for prediction accuracy and distributional calibration. Our systematic comparison of LLM approaches, model architectures, training approaches, and data scaling reveals that Mistral-7B consistently outperforms encoder architectures, embedding-based methods, and few-shot learning methods. Our experiments also reveal the effectiveness of LLM-assisted label correction in achieving human-level accuracy without systematic bias. We will release our curated datasets with standardized cleaning procedures to support future research in this direction.

LAM SIMULATOR: Advancing Data Generation for Large Action Models Trainings via Online Exploration and Feedback Simulation

Thai Quoc Hoang, Kung-Hsiang Huang, Shirley Kokane, Jianguo Zhang, Zuxin Liu, Ming Zhu, Jake Grigsby, Tian Lan, Michael S Ryoo, Chien-Sheng Wu, Shelby Heinecke, Huan Wang, Silvio Savarese, Caiming Xiong, Juan Carlos Niebles

Large Action Models (LAMs) for AI Agents offer incredible potential but face challenges due to the need for high-quality training data, especially for multi-steps tasks that involve planning, executing tool calls, and responding to feedback. To address these issues, we present LAM SIMULATOR, a comprehensive framework designed for online exploration of agentic tasks with high-quality feedback. Our framework features a dynamic task query generator, an extensive collection of tools, and an interactive environment where Large Language Model (LLM) Agents can call tools and receive real-time feedback. This setup enables LLM Agents to explore and solve tasks independently and potentially come up with multiple approaches to tackle any given task. Generated data are then used to create high-quality training datasets for LAMs. Our research shows that LAM SIMULATOR enables LLM Agents to autonomously solve tasks while automating the creation of high-quality training data. Models trained with these self-generated datasets demonstrated significant performance gains, showing up to a 49.3% improvement over their own baselines. This was especially evident in experiments conducted with the ToolBench and CRMArena environments. The process requires minimal human input during dataset creation, highlighting the LAM SIMULATOR's efficiency and effectiveness in speeding up AI agents' development.

Lifelong Model Editing with Graph-Based External Memory

Yash Kumar Atri, Ahmed Alaa, Thomas Hartvigsen

Large language models (LLMs) have revolutionized natural language processing, yet their practical utility is often limited by persistent issues of hallucinations and outdated parametric knowledge. Although post-training model editing offers a pathway for dynamic updates, existing methods frequently suffer from overfitting and catastrophic forgetting. To tackle these challenges, we propose a novel framework that leverages hyperbolic geometry and graph neural networks for precise and stable model edits. We introduce HYPE, (HYperbolic Parameter Editing), which comprises three key components: (i) Hyperbolic Graph Construction, which uses Poincaré embeddings to represent knowledge triples in hyperbolic space, preserving hierarchical relationships and preventing unintended side effects by ensuring that edits to parent concepts do not inadvertently affect child concepts; (ii) Möbius-Transformed Updates, which apply hyperbolic addition to propagate edits while maintaining structural consistency within the hyperbolic manifold, unlike conventional Euclidean updates that distort relational distances; and (iii) Dual Stabilization, which combines gradient masking and periodic GNN parameter resetting to prevent catastrophic forgetting by focusing updates on critical parameters and preserving long-term knowledge. Experiments on CounterFact, CounterFact+, and MQuAKE with GPT-J and GPT2-XL demonstrate that HYPE significantly enhances edit stability, factual accuracy, and multi-hop reasoning.

When Detection Fails: The Power of Fine-Tuned Models to Generate Human-Like Social Media Text

Hillary Dawkins, Kathleen C. Fraser, Svetlana Kiritchenko

Detecting AI-generated text is a difficult problem to begin with; detecting AI-generated text on social media is made even more difficult due to the short text length and informal, idiosyncratic language of the internet. It is nonetheless important to tackle this problem, as social media represents a significant attack vector in online influence campaigns, which may be bolstered through the use of mass-produced AI-generated posts supporting (or opposing) particular policies, decisions, or events. We approach this problem with the mindset and resources of a reasonably sophisticated threat actor, and create a dataset of 505,159 AI-generated social media posts from a combination of open-source, closed-source, and fine-tuned LLMs, covering 11 different controversial topics. We show that while the posts can be detected under typical research assumptions about knowledge of and access to the generating models, under the more realistic assumption that an attacker will not release their fine-tuned model to the public, detectability drops dramatically. This result is confirmed with a human study. Ablation experiments highlight the vulnerability of various detection algorithms to fine-tuned LLMs. This result has implications across all detection domains, since fine-tuning is a generally applicable and realistic LLM use case.

Towards Medical Complex Reasoning with LLMs through Medical Verifiable Problems

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Benyou Wang

The breakthrough of OpenAI o1 highlights the potential of enhancing reasoning to improve LLM. Yet, most research in reasoning has focused on mathematical tasks, leaving domains like medicine underexplored. The medical domain, though distinct from mathematics, also demands robust reasoning to provide reliable answers, given the high standards of healthcare. However, verifying medical reasoning is challenging, unlike those in mathematics. To address this, we propose **Medical Verifiable Problems** with a medical verifier to check the correctness of model outputs. This verifiable nature enables advancements in medical reasoning through **a two-stage approach**: (1) using the verifier to guide the search for a complex reasoning trajectory for fine-tuning LLMs, (2) applying reinforcement learning (RL) with verifier-based rewards to enhance complex reasoning further. Finally, we introduce HuatuoGPT-o1, a medical LLM capable of complex reasoning, which outperforms general and medical-specific baselines using only 40K verifiable problems. Experiments show complex reasoning improves medical problem-solving and benefits more from RL. We hope our approach inspires advancements in reasoning across medical and other specialized domains. Code, datasets, and models are publicly available at https://github.com/FreedomIntelligence/HuatuoGPT-o1.

LLM Critics Help Catch Bugs in Mathematics: Towards a Better Mathematical Verifier with Natural Language Feedback

Bofei Gao, Zefan Cai, Runxin Xu, Peiyi Wang, Ce Zheng, Runji Lin, Keming Lu, Dayiheng Liu, Chang Zhou, Wen Xiao, Tianyu Liu, Baobao Chang

In recent progress, mathematical verifiers have achieved success in mathematical reasoning tasks by validating the correctness of solutions generated by policy models. However, existing verifiers are trained with binary classification labels, which are not informative enough for the model to accurately assess the solutions. To mitigate the aforementioned insufficiency of binary labels, we introduce step-wise natural language feedback as rationale labels, that is, the correctness of each step and the detailed explanations. In this paper, we propose Math-Minos, a natural language feedback-enhanced verifier by constructing automatically generated training data and a two-stage training paradigm for effective training and efficient inference. Our experiments reveal that a small set of natural language feedback can significantly boost the performance of the verifier in both verification and reinforcement learning and also significantly alleviates the data-demanding problems of the reward model with an over 700% data efficiency improvement.

Exploring LLM Annotation for Adaptation of Clinical Information Extraction Models under Data-sharing Restrictions

Seiji Shimizu, HISADA Shohei, Yutaka Uno, Shuntaro Yada, Shoko Wakamiya, Eiji Aramaki

In-hospital text data contains valuable clinical information, yet deploying fine-tuned small language models (SLMs) for information extraction remains challenging due to differences in formatting and vocabulary across institutions. Since access to the original in-hospital data (source domain) is often restricted, annotated data from the target hospital (target domain) is crucial for domain adaptation. However, clinical annotation is notoriously expensive and time-consuming, as it demands clinical and linguistic expertise. To address this issue, we leverage large language models (LLMs) to annotate the target domain data for the adaptation. We conduct experiments on four clinical information extraction tasks, including eight target domain data. Experimental results show that LLM-annotated data consistently enhances SLM performance and, with a larger number of annotated data, outperforms manual annotation in three out of four tasks.

RecordTwin: Towards Creating Safe Synthetic Clinical Corpora

Seiji Shimizu, Ibrahim Baroud, Lisa Raithel, Shuntaro Yada, Shoko Wakamiya, Eiji Aramaki

The scarcity of publicly available clinical corpora hinders developing and applying NLP tools in clinical research. While existing work tackles this issue by utilizing generative models to create high-quality synthetic corpora, their methods require learning from the original in-hospital clinical documents, turning them unfeasible in practice. To address this problem, we introduce RecordTwin, a novel synthetic corpus creation method designed to generate synthetic documents from anonymized clinical entities. In this method, we first extract and anonymize entities from in-hospital documents to ensure the information contained in the synthetic corpus is restricted. Then, we use a large language model to fill the context between anonymized entities. To do so, we use a small, privacy-preserving subset of the original documents to mimic their formatting and writing style. This approach only requires anonymized entities and a small subset of original documents in the generation process, making it more feasible in practice. To evaluate the synthetic corpus created with our method, we conduct a proof-of-concept study using a publicly available clinical database. Our results demonstrate that the synthetic corpus has a utility comparable to the original data and a safety advantage over baselines, highlighting the potential of RecordTwin for privacy-preserving synthetic corpus creation.

PGPO: Enhancing Agent Reasoning via Pseudocode-style Planning Guided Preference Optimization

zouying cao, Runze Wang, Yifei Yang, Xinbei Ma, Xiaoyong Zhu, Bo Zheng, hai zhao

Large Language Model (LLM) agents have demonstrated impressive capabilities in handling complex interactive problems. Existing LLM agents mainly generate natural language plans to guide reasoning, which is verbose and inefficient. NL plans are also tailored to specific tasks and restrict agents' ability to generalize across similar tasks. To this end, we explore pseudocode-style plans (P-code Plan) to capture the structural logic of reasoning. We find that P-code Plan empowers LLM agents with stronger generalization ability and more efficiency. Inspired by this finding, we propose a pseudocode-style \underline{P}lanning \underline{G}uided \underline{P}reference \underline{O}ptimization method called PGPO for effective agent learning. With two planning-oriented rewards, PGPO further enhances LLM agents' ability to generate high-quality p-code plans and subsequent reasoning. Experiments show that PGPO achieves superior performance on representative agent benchmarks and outperforms the current leading baselines. Analyses reveal the advantage of PGPO in reducing action errors and omissions during reasoning.

Grammar-Based Code Representation: Is It a Worthy Pursuit for LLMs?

Qingyuan Liang, Zhao Zhang, Zeyu Sun, Zheng Lin, Qi Luo, Xiao Yueyi, Yizhou Chen, Yuqun Zhang, Haotian Zhang, Lu Zhang, chenbin, Yingfei Xiong

Grammar serves as a cornerstone in programming languages and software engineering, providing frameworks to define the syntactic space and program structure. Existing research demonstrates the effectiveness of grammar-based code representations in small-scale models, showing their ability to reduce syntax errors and enhance performance. However, as language models scale to the billion level or beyond, syntax-level errors become rare, making it unclear whether grammar information still provides performance benefits. To explore this, we develop a series of billion-scale GrammarCoder models, incorporating grammar rules in the code generation process. Experiments on HumanEval (+) and MBPP (+) demonstrate a notable improvement in code generation accuracy. Further analysis shows that grammar-based representations enhance LLMs' ability to discern subtle code differences, reducing semantic errors caused by minor variations. These findings suggest that grammar-based code representations remain valuable even in billion-scale models, not only by maintaining syntax correctness but also by improving semantic differentiation.

Token-level Preference Self-Alignment Optimization for Multi-style Outline Controllable Generation

Zihao Li, Xuekong Xu, Ziyao Chen, Lixin Zou, Ethanhjwu, Qiang Chen, Chenliang Li

Multi-style outline controllable generation is crucial for multiple applications, including document semantic structuring and retrieval-augmented generation. The great success of preference alignment approaches encourages their application in controllable generation tasks. However, these attempts encounter several limitations: (1) response pair requirements, (2) substantial computation costs, and (3) insufficient exploitation of fine-grained preference signals. To address these problems, we propose a token-level preference self-alignment optimization, named TKPO, for outline controllable generation. TKPO extends the Bradley-Terry model from pair-wise to list-wise comparison, which is further applied at the token level for fine-grained preference signal utilization. In comparison to the representative methods, e.g., DPO, TKPO does not require response pairs; instead, we propose a controllable attributes-driven method to construct reject samples for self-alignment. Additionally, TKPO optimizes only the base model, thereby avoiding additional memory usage and substantial computational costs. We curate two outline controllable generation datasets with regard to language style and level-of-detail. Extensive experiments demonstrate that TKPO outperforms DPO by up to 19.28% in performance while requiring only 56.25% in training time. We release the code and datasets resources at https://github.com/WHUIR/TKPO.

Evaluating Large Language Models for Confidence-based Check Set Selection

Jane Arleth dela Cruz, Iris Hendrickx, Martha Larson

Large Language Models (LLMs) have shown promise in automating high-labor data tasks, but the adoption of LLMs in high-stake scenarios faces two key challenges: their tendency to answer despite uncertainty and their difficulty handling long input contexts robustly. We investigate commonly used off-the-shelf LLMs' ability to identify low-confidence outputs for human review through "check set selection"--a process where LLMs prioritize information needing human judgment. Using a case study on social media monitoring for disaster risk management, we define the “check set” as a list of tweets escalated to the disaster manager when the LLM has the least confidence, enabling human oversight within budgeted effort. We test two strategies for LLM check set selection: *individual confidence elicitation* -- LLMs assesses confidence for each tweet classification individually, requiring more prompts with shorter contexts, and *direct set confidence elicitation* -- LLM evaluates confidence for a list of tweet classifications at once, using less prompts but longer contexts. Our results reveal that set selection via individual probabilities is more reliable but that direct set confidence merits further investigation. Direct set selection challenges include inconsistent outputs, incorrect check set size, and low inter-annotator agreement. Despite these challenges, our approach improves collaborative disaster tweet classification by outperforming random-sample check set selection, demonstrating the potential of human-LLM collaboration.

EMRs2CSP : Mining Clinical Status Pathway from Electronic Medical Records

Yifei Chen, Ruihui Hou, Jingping Liu, Tong Ruan

Many current studies focus on extracting tests or treatments when constructing clinical pathways, often neglecting the patient's symptoms and diagnosis, leading to incomplete diagnostic and therapeutic logic. Therefore, this paper aims to extract clinical pathways from electronic medical records that encompass complete diagnostic and therapeutic logic, including temporal information, patient symptoms, diagnosis, and tests or treatments. To achieve this objective, we propose a novel clinical pathway representation: the clinical status pathway. We also design a LLM-based pipeline framework for extracting clinical status pathway from electronic medical records, with the core concept being to improve extraction accuracy by modeling the diagnostic and treatment processes. In our experiments, we apply this framework to construct a comprehensive breast cancer-specific clinical status pathway and evaluate its performance on medical question-answering and decision-support tasks, demonstrating significant improvements over traditional clinical pathways. The code is publicly available at https://anonymous.4open.science/r/EMR2CSP-BD11

Hatevolution: What Static Benchmarks Don't Tell Us

Chiara Di Bonaventura, Barbara McGillivray, Yulan He, Albert Meroño-Peñuela

Language changes over time, including in the hate speech domain, which evolves quickly following social dynamics and cultural shifts. While NLP research has investigated the impact of language evolution on model training and has proposed several solutions for it, its impact on model benchmarking remains under-explored. Yet, hate speech benchmarks play a crucial role to ensure model safety. In this paper, we empirically evaluate the robustness of 20 language models across two evolving hate speech experiments, and we show the temporal misalignment between static and time-sensitive evaluations. Our findings call for time-sensitive evaluation benchmarks in order to correctly and reliably evaluate language models.

Tag-Instruct: Controlled Instruction Complexity Enhancement through Structure-based Augmentation

He Zhu, Zhiwen Ruan, Junyou Su, Xingwei He, Wenjia Zhang, Yun Chen, Guanhua Chen

High-quality instruction data is crucial for developing large language models (LLMs), yet existing approaches struggle to effectively control instruction complexity. We present Tag-Instruct, a novel framework that enhances instruction complexity through structured semantic compression and controlled difficulty augmentation. Unlike previous prompt-based methods operating on raw text, Tag-Instruct compresses instructions into a compact tag space and systematically enhances complexity through RL-guided tag expansion. Through extensive experiments, we show that Tag-Instruct outperforms existing instruction complexity augmentation approaches. Our analysis reveals that operating in tag space provides superior controllability and stability across different instruction synthesis frameworks.

Exploring Anthropomorphic Language in the Reporting of NLP Findings

Matthew Shardlow, Ashley Williams, Charlie Roadhouse, Filippos Ventirozos, Piotr Przybyła

We investigate the prevalence of anthropomorphic language in the reporting of AI technology, focussed on NLP and LLMs. We undertake a corpus annotation focussing on one year of ACL long-paper abstracts and news articles from the same period. We find that 74% of ACL abstracts and 88% of news articles contain some form of anthropomorphic description of AI technology. Further, we train a regression classifier based on BERT, demonstrating that we can automatically label abstracts for their degree of anthropomorphism based on our corpus. We conclude by applying this labelling process to abstracts available in the entire history of the ACL Anthology and reporting on diachronic and inter-venue findings, showing that the degree of anthropomorphism is increasing at all examined venues over time.

Optimizing Decomposition for Optimal Claim Verification

Yining Lu, Noah Ziems, Hy Dang, Meng Jiang

Current research on the Decompose-Then-Verify paradigm for evaluating the factuality of long-form text typically treats decomposition and verification in isolation, overlooking their interactions and potential misalignment. We find that existing decomposition policies, typically hand-crafted demonstrations, do not align well with downstream verifiers in terms of atomicity—a novel metric quantifying information density—leading to suboptimal verification results. We formulate finding the optimal decomposition policy for optimal verification as a bilevel optimization problem. To approximate a solution for this strongly NP-hard problem, we propose dynamic decomposition, a reinforcement learning framework that leverages verifier feedback to learn a policy for dynamically decomposing claims to verifier-preferred atomicity. Experimental results show that dynamic decomposition outperforms existing decomposition policies, improving verification confidence by 0.07 and accuracy by 0.12 (on a 0-1 scale) on average across varying verifiers, datasets, and atomcities of input claims.

Enhancing Transformers for Generalizable First-Order Logical Entailment

Tianshi Zheng, Jiazheng Wang, Zihao Wang, Jiaxin Bai, Hang Yin, Zheye Deng, Yangqiu Song, Jianxin Li

Transformers, as the fundamental deep learning architecture, have demonstrated great capability in reasoning. This paper studies the generalizable first-order logical reasoning ability of transformers with their *parameterized* knowledge and how to improve it. Transformers' capability of first-order reasoning is further captured by whether they can conduct first-order logical entailment, which is quantitatively measured by their performance in answering knowledge graph queries. We establish the connections between (1) two types of distribution shifts studied in out-of-distribution generalization and (2) unseen knowledge and query settings discussed in the task of knowledge graph query answering, which makes it possible to characterize the fine-grained generalizability. Results on our comprehensive dataset showed that transformers **outperform** previous methods designed particularly for this task and provided detailed empirical evidence about the impact of the input query syntax, token embedding, and transformer architectures on the reasoning capability of transformers. Interestingly, our results revealed the mismatch of positional encoding and other design choices of transformer architectures in previous practices. Motivated by this, we propose **TEGA**, a logic-aware architecture that significantly improves the performance in generalizable first-order logical entailment.

MathFusion: Enhancing Mathematical Problem-solving of LLM through Instruction Fusion

Qizhi Pei, Lijun Wu, Zhuoshi Pan, Yu Li, Honglin Lin, Chenlin Ming, Xin Gao, Conghui He, Rui Yan

Large Language Models (LLMs) have shown impressive progress in mathematical reasoning. While data augmentation is promising to enhance mathematical problem-solving ability, current approaches are predominantly limited to instance-level modifications—such as rephrasing or generating syntactic variations—which fail to capture and leverage the intrinsic relational structures inherent in mathematical knowledge. Inspired by human learning processes, where mathematical proficiency develops through systematic exposure to interconnected concepts, we introduce \textbf{\mbox{MathFusion}}, a novel framework that enhances mathematical reasoning through cross-problem instruction synthesis. MathFusion implements this through three fusion strategies: (1) \textit{sequential fusion}, which chains related problems to model solution dependencies; (2) \textit{parallel fusion}, which combines analogous problems to reinforce conceptual understanding; and (3) \textit{conditional fusion}, which creates context-aware selective problems to enhance reasoning flexibility. By applying these strategies, we generate a new dataset, \textbf{MathFusionQA}, followed by fine-tuning models (DeepSeekMath-7B, Mistral-7B, Llama3-8B) on it. Experimental results demonstrate that MathFusion achieves substantial improvements in mathematical reasoning while maintaining high data efficiency, boosting performance by 18.0 points in accuracy across diverse benchmarks while requiring only 45K additional synthetic instructions, representing a substantial improvement over traditional single-instruction approaches.

Aligning Large Language Models with Implicit Preferences from User-Generated Content

Zhaoxuan Tan, Zheng Li, Tianyi Liu, Haodong Wang, Hyokun Yun, Ming Zeng, Pei Chen, Zhihan Zhang, Yifan Gao, Ruijie Wang, Priyanka Nigam, Bing Yin, Meng Jiang

Learning from preference feedback is essential for aligning large language models (LLMs) with human values and improving the quality of generated responses. However, existing preference learning methods rely heavily on curated data from humans or advanced LLMs, which is costly and difficult to scale. In this work, we present PUGC, a novel framework that leverages implicit human Preferences in unlabeled User-Generated Content (UGC) to generate preference data. Although UGC is not explicitly created to guide LLMs in generating human-preferred responses, it often reflects valuable insights and implicit preferences from its creators that has the potential to address readers' questions. PUGC transforms UGC into user queries and generates responses from the policy model. The UGC is then leveraged as a reference text for response scoring, aligning the model with these implicit preferences. This approach improves the quality of preference data while enabling scalable, domain-specific alignment. Experimental results on Alpaca Eval 2 show that models trained with DPO and PUGC achieve a 9.37% performance improvement over traditional methods, setting a 35.93% state-of-the-art length-controlled win rate using Mistral-7B-Instruct. Further studies highlight gains in reward quality, domain-specific alignment effectiveness, robustness against UGC quality, and theory of mind capabilities. Our code and dataset are available at https://zhaoxuan.info/PUGC.github.io/.

PrivaCI-Bench: Evaluating Privacy with Contextual Integrity and Legal Compliance

Haoran Li, Wenbin Hu, Huihao JING, Yulin Chen, Qi Hu, Sirui Han, Tianshu Chu, Peizhao Hu, Yangqiu Song

Recent advancements in generative large language models (LLMs) have enabled wider applicability, accessibility, and flexibility. However, their reliability and trustworthiness are still in doubt, especially for concerns regarding individuals' data privacy. Great efforts have been made on privacy by building various evaluation benchmarks to study LLMs' privacy awareness and robustness from their generated outputs to their hidden representations. Unfortunately, most of these works adopt a narrow formulation of privacy and only investigate personally identifiable information (PII). In this paper, we follow the merit of the Contextual Integrity (CI) theory, which posits that privacy evaluation should not only cover the transmitted attributes but also encompass the whole relevant social context through private information flows. We present PrivaCI-Bench, a comprehensive contextual privacy evaluation benchmark targeted at legal compliance to cover well-annotated privacy and safety regulations, real court cases, privacy policies, and synthetic data built from the official toolkit to study LLMs' privacy and safety compliance. We evaluate the latest LLMs, including the recent reasoner models QwQ-32B and Deepseek R1. Our experimental results suggest that though LLMs can effectively capture key CI parameters inside a given context, they still require further advancements for privacy compliance.

PrivacyRestore: Privacy-Preserving Inference in Large Language Models via Privacy Removal and Restoration

Ziqian Zeng, Jianwei Wang, Junyao Yang, ZhengdongLu, Haoran Li, Huiping Zhuang, Cen Chen

The widespread usage of online Large Language Models (LLMs) inference services has raised significant privacy concerns about the potential exposure of private information in user inputs. Existing privacy protection methods for LLMs suffer from either insufficient privacy protection with performance degradation, or large inference time overhead. To address these limitations, we propose PrivacyRestore, a plug-and-play method to protect the privacy of user inputs during LLM inference for the client-server scenario. The server first trains restoration vectors for each privacy span type offline and then releases them to the clients. During inference, the client aggregates restoration vectors of all privacy spans in the user query into a meta restoration vector, which is later sent to the server to restore information. Before transmission, the client removes all privacy spans in the user query and applies d_\chi-privacy mechanism to the meta vector for privacy protection. We prove that our method can inherently prevent the linear growth of the privacy budget. We conduct extensive experimental, covering the medical and legal domains, and demonstrate that PrivacyRestore effectively protects private information and maintains acceptable levels of performance and inference efficiency

Large Language and Protein Assistant for Protein-Protein Interactions Prediction

Peng Zhou, Pengsen Ma, Jianmin Wang, Xibao Cai, Haitao Huang, Wei Liu, Longyue Wang, Lai Hou Tim, xiangxiang Zeng

Predicting the types and affinities of protein-protein interactions (PPIs) is crucial for understanding biological processes and developing novel therapeutic approaches. While encoding proteins themselves is essential, PPI networks can also provide rich prior knowledge for these predictive tasks. However, existing methods oversimplify the problem of PPI prediction in a semi-supervised manner when utilizing PPI networks, limiting their practical application. Furthermore, how to effectively use the rich prior knowledge of PPI networks for novel proteins not present in the network remains an unexplored issue. Additionally, due to inflexible architectures, most of existing methods cannot handle complexes containing an flexible number of proteins. To overcome these limitations, we introduce LLaPA (Large Language and Protein Assistant), a multimodal large language model that integrates proteins and PPI networks. LLaPA offers a more rational approach to utilizing PPI networks for PPI prediction and can fully exploit the information of PPI networks for unseen proteins. Through natural language instructions, LLaPA can accept flexible number of protein sequences and has the potential to perform various protein tasks. Experiments show that LLaPA achieves state-of-the-art performance in multi-label PPI (mPPI) type prediction and is capable of predicting the binding affinity between multiple interacting proteins based on sequence data.

GuideBench: Benchmarking Domain-Oriented Guideline Following for LLM Agents

Lingxiao Diao, Xinyue Xu, Wanxuan Sun, Cheng Yang, Zhuosheng Zhang

Large language models (LLMs) have been widely deployed as autonomous agents capable of following user instructions and making decisions in real-world applications. Previous studies have made notable progress in benchmarking the instruction following capabilities of LLMs in general domains, with a primary focus on their inherent commonsense knowledge. Recently, LLMs have been increasingly deployed as domain-oriented agents, which rely on domain-oriented guidelines that may conflict with their commonsense knowledge. These guidelines exhibit two key characteristics: they consist of a wide range of domain-oriented rules and are subject to frequent updates. Despite these challenges, the absence of comprehensive benchmarks for evaluating the domain-oriented guideline following capabilities of LLMs presents a significant obstacle to their effective assessment and further development. In this paper, we introduce GuideBench, a comprehensive benchmark designed to evaluate guideline following performance of LLMs. GuideBench evaluates LLMs on three critical aspects: (i) adherence to diverse rules, (ii) robustness to rule updates, and (iii) alignment with human preferences. Experimental results on a range of LLMs indicate substantial opportunities for improving their ability to follow domain-oriented guidelines. Data and code are available at Anonymous.

PwnGPT: Automatic Exploit Generation Based on Large Language Models

Wanzong Peng, Lin Ye, Xuetao Du, Hongli Zhang, Dongyang Zhan, Yunting Zhang, Yicheng Guo, Chen Zhang

Automatic exploit generation (AEG) refers to the automatic discovery and exploitation of vulnerabilities against unknown targets. Traditional AEG often targets a single type of vulnerability and still relies on templates built from expert experience. To achieve intelligent exploit generation, we establish a comprehensive benchmark using Binary Exploitation (pwn) challenges in Capture the Flag (CTF) competitions and investigate the capabilities of Large Language Models (LLMs) in AEG based on the benchmark. To improve the performance of AEG, we propose PwnGPT, an LLM-based automatic exploit generation framework that automatically solves pwn challenges. The structural design of PwnGPT is divided into three main components: analysis, generation, and verification modules. With the help of a modular approach and structured problem inputs, PwnGPT can solve challenges that LLMs cannot directly solve. We evaluate PwnGPT on our benchmark and analyze the outputs of each module. Experimental results show that our framework is highly autonomous and capable of addressing various challenges. Compared to direct input LLMs, PwnGPT increases the completion rate of exploit on our benchmark from 26.3\% to 57.9\% with the OpenAI o1-preview model and from 21.1\% to 36.8\% with the GPT-4o model.

PaSa: An LLM Agent for Comprehensive Academic Paper Search

Yichen He, Guanhua Huang, Peiyuan Feng, Yuan Lin, Yuchen Zhang, Hang Li, Weinan E

We introduce PaSa, an advanced Paper Search agent powered by large language models. PaSa can autonomously make a series of decisions, including invoking search tools, reading papers, and selecting relevant references, to ultimately obtain comprehensive and accurate results for complex scholar queries. We optimize PaSa using reinforcement learning with a synthetic dataset, AutoScholarQuery, which includes 35k fine-grained academic queries and corresponding papers sourced from top-tier AI conference publications. Additionally, we develop RealScholarQuery, a benchmark collecting real-world academic queries to assess PaSa performance in more realistic scenarios. Despite being trained on synthetic data, PaSa significantly outperforms existing baselines on RealScholarQuery, including Google, Google Scholar, Google with GPT-4o for paraphrased queries, ChatGPT (search-enabled GPT-4o), GPT-o1, and PaSa-GPT-4o (PaSa implemented by prompting GPT-4o). Notably, PaSa-7B surpasses the best Google-based baseline, Google with GPT-4o, by 37.78% in recall@20 and 39.90% in recall@50, and exceeds PaSa-GPT-4o by 30.36% in recall and 4.25% in precision. Model, datasets, and code are available at https://github.com/bytedance/pasa. Demo: https://pasa-agent.ai

DebateCoder: Towards Collective Intelligence of LLMs via Test Case Driven LLM Debate for Code Generation

Jizheng Chen, Kounianhua Du, Xinyi Dai, Weiming Zhang, Xihuai Wang, Yasheng Wang, Ruiming Tang, Weinan Zhang, Yong Yu

With the impressive reasoning and text generation capabilities of large language models (LLMs), methods leveraging multiple LLMs to debate each other have garnered increasing attention. However, existing debate-based approaches remain limited in effectiveness in structured and detailed domains represented by code generation due to several reasons: 1) Reliance on different instances of the same LLM for debate, neglecting the potential benefits of integrating diverse models with varied internal knowledge for more comprehensive code generation, 2) under-utilization of test cases, and 3) reliance on third-party LLM moderators for result consolidation and decision-making, probably introducing hallucinations and judgment errors. To address these challenges, we propose DebateCoder to collect intelligence of LLMs via test case-driven debate for code generation. In DebateCoder, test cases serve as a medium for models to analyze code and identify bugs, while opposing models generate test cases to challenge each other's code during the debate process. These test cases, along with their execution results, are elaborately leveraged to refine and enhance the code through a novel contrastive analysis process. Furthermore, DebateCoder leverages test case outcomes to assess code quality and determine convergence criteria. Unlike previous approaches, DebateCoder emphasizes the collaborative improvement of both models through competitive debate and interactive analysis. Abundant experimental results on two datasets demonstrate the effectiveness of DebateCoder.

Can a Single Model Master Both Multi-turn Conversations and Tool Use? CALM: A Unified Conversational Agentic Language Model

Emre Can Acikgoz, Jeremiah Greer, Akul Datta, Ze Yang, William Zeng, Oussama Elachqar, Emmanouil Koukoumidis, Dilek Hakkani-Tür, Gokhan Tur

Large Language Models (LLMs) with API-calling capabilities enabled building effective Language Agents (LA), while also revolutionizing the conventional task-oriented dialogue (TOD) paradigm. However, current approaches face a critical dilemma: TOD systems are often trained on a limited set of target APIs, requiring new data to maintain their quality when interfacing with new services, while LAs are not trained to maintain user intent over multi-turn conversations. Because both robust multi-turn management and advanced function calling are crucial for effective conversational agents, we evaluate these skills on three popular benchmarks: MultiWOZ 2.4 (TOD), BFCL V3 (LA), and API-Bank (LA)—and our analyses reveal that specialized approaches excel in one domain but underperform in the other. To bridge this chasm, we introduce **CoALM** (**C**onversational **A**gentic **L**anguage **M**odel), a unified approach that integrates both conversational and agentic capabilities. We created **CoALM-IT**, a carefully constructed multi-task dataset that interleave multi-turn ReAct reasoning with complex API usage. Using CoALM-IT, we train three models **CoALM 8B**, **CoALM 70B**, and **CoALM 405B**, which outperform top domain-specific models, including GPT-4o, across all three benchmarks. This demonstrates the feasibility of a single model approach for both TOD and LA, setting a new standard for conversational agents.

SINCon: Mitigate LLM-Generated Malicious Message Injection Attack for Rumor Detection

Mingqing Zhang, Qiang Liu, Xiang Tao, Shu Wu, Liang Wang

In the era of rapidly evolving large language models (LLMs), state-of-the-art rumor detection systems, particularly those based on Message Propagation Trees (MPTs), which represent a conversation tree with the post as its root and the replies as its descendants, are facing increasing threats from adversarial attacks that leverage LLMs to generate and inject malicious messages. Existing methods are based on the assumption that different nodes exhibit varying degrees of influence on predictions. They define nodes with high predictive influence as important nodes and target them for attacks. If the model treats nodes' predictive influence more uniformly, attackers will find it harder to target high predictive influence nodes. In this paper, we propose Similarizing the predictive Influence of Nodes with Contrastive Learning (SINCon), a defense mechanism that encourages the model to learn graph representations where nodes with varying importance have a more uniform influence on predictions. Extensive experiments on the Twitter and Weibo datasets demonstrate that SINCon not only preserves high classification accuracy on clean data but also significantly enhances resistance against LLM-driven message injection attacks.

Improving Medical Large Vision-Language Models with Abnormal-Aware Feedback

Yucheng Zhou, Lingran Song, Jianbing Shen

Existing Medical Large Vision-Language Models (Med-LVLMs), encapsulating extensive medical knowledge, demonstrate excellent capabilities in understanding medical images. However, there remain challenges in visual localization in medical images, which is crucial for abnormality detection and interpretation. To address these issues, we propose a novel UMed-LVLM designed to unveil medical abnormalities. Specifically, we collect a Medical Abnormalities Unveiling (MAU) dataset and propose a two-stage training method for UMed-LVLM training. To collect MAU dataset, we propose a prompt method utilizing the GPT-4V to generate diagnoses based on identified abnormal areas in medical images. Moreover, the two-stage training method includes Abnormal-Aware Instruction Tuning and Abnormal-Aware Rewarding, comprising Relevance Reward, Abnormal Localization Reward and Vision Relevance Reward. Experimental results demonstrate that our UMed-LVLM significantly outperforms existing Med-LVLMs in identifying and understanding medical abnormalities, achieving a 58% improvement over the baseline. In addition, this work shows that enhancing the abnormality detection capabilities of Med-LVLMs significantly improves their understanding of medical images and generalization capability. Our code and data release at URL.

Genius: A Generalizable and Purely Unsupervised Self-Training Framework For Advanced Reasoning

Fangzhi Xu, Hang Yan, Chang Ma, Haiteng Zhao, Qiushi Sun, Kanzhi Cheng, Junxian He, Jun Liu, Zhiyong Wu

Advancing LLM reasoning skills has captivated wide interest. However, current post-training techniques rely heavily on supervisory signals, such as outcome supervision or auxiliary reward models, which face the problem of scalability and high annotation costs. This motivates us to enhance LLM reasoning without the need for external supervision. Given the input query, the LLM seeks the globally optimal response by stepwise sampling and self-rewarding, and optimizes itself with the collected responses. Genius offers some technical solutions to address the following key challenges. To tackle the problem of how to determine the steps in the response via self-rewarding, Genius introduces a stepwise foresight re-sampling strategy to sample and estimate the step value by simulating future outcomes. Recognizing the intrinsic noise and uncertainty of self-supervision, we propose an advantage-calibrated optimization (ACO) loss function to mitigate estimation inconsistencies. In short, Genius provides an advanced initial step towards self-improve LLM reasoning with general queries and without supervision, revolutionizing reasoning scaling laws given the vast availability of general queries.

Continual Gradient Low-Rank Projection Fine-Tuning for LLMs

Chenxu Wang, Yilin Lyu, Zicheng Sun, Liping Jing

Continual fine-tuning of Large Language Models (LLMs) is hampered by the trade-off between efficiency and expressiveness. Low-Rank Adaptation (LoRA) offers efficiency but constrains the model's ability to learn new tasks and transfer knowledge due to its low-rank nature and reliance on explicit parameter constraints. We propose GORP (\underline{\textbf{G}}radient L\underline{\textbf{O}}w \underline{\textbf{R}}ank \underline{\textbf{P}}rojection) for Continual Learning, a novel training strategy that overcomes these limitations by synergistically combining full and low-rank parameters and jointly updating within a unified low-rank gradient subspace. GORP expands the optimization space while preserving efficiency and mitigating catastrophic forgetting. Extensive experiments on continual learning benchmarks demonstrate GORP's superior performance compared to existing state-of-the-art approaches. Code is available at https://github.com/Wcxwcxw/GORP.

GraphCheck: Breaking Long-Term Text Barriers with Extracted Knowledge Graph-Powered Fact-Checking

Yingjian Chen, Haoran Liu, Yinhong Liu, Jinxiang Xie, Rui Yang, Han Yuan, Yanran Fu, Peng Yuan Zhou, Qingyu Chen, James Caverlee, Irene Li

Large language models (LLMs) are widely used, but they often generate subtle factual errors, especially in long-form text. These errors are fatal in some specialized domains such as medicine. Existing fact-checking with grounding documents methods face two main challenges: (1) they struggle to understand complex multihop relations in long documents, often overlooking subtle factual errors; (2) most specialized methods rely on pairwise comparisons, requiring multiple model calls, leading to high resource and computational costs. To address these challenges, we propose GraphCheck, a fact-checking framework that uses extracted knowledge graphs to enhance text representation. Graph Neural Networks further process these graphs as a soft prompt, enabling LLMs to incorporate structured knowledge more effectively. Enhanced with graph-based reasoning, GraphCheck captures multihop reasoning chains that are often overlooked by existing methods, enabling precise and efficient fact-checking in a single inference call. Experimental results on seven benchmarks spanning both general and medical domains demonstrate up to a 7.1\% overall improvement over baseline models. Notably, GraphCheck outperforms existing specialized fact-checkers and achieves comparable performance with state-of-the-art LLMs, such as DeepSeek-V3 and OpenAI-o1, with significantly fewer parameters.

PreP-OCR: A Complete Pipeline for Document Image Restoration and Enhanced OCR Accuracy

Shuhao Guan, Moule Lin, Cheng Xu, Xinyi Liu, Jinman Zhao, Jiexin Fan, Qi Xu, Derek Greene

This paper introduces PreP-OCR, a two-stage pipeline that combines document image restoration with semantic-aware post-OCR correction to enhance both visual clarity and textual consistency, thereby improving text extraction from degraded historical documents. First, we synthesize document-image pairs from plaintext, rendering them with diverse fonts and layouts and then applying a randomly ordered set of degradation operations. An image restoration model is trained on this synthetic data, using multi-directional patch extraction and fusion to process large images. Second, a ByT5 post-OCR model, fine-tuned on synthetic historical text pairs, addresses remaining OCR errors. Detailed experiments on 13,831 pages of real historical documents in English, French, and Spanish show that the PreP-OCR pipeline reduces character error rates by 63.9-70.3\% compared to OCR on raw images. Our pipeline demonstrates the potential of integrating image restoration with linguistic error correction for digitizing historical archives.

IRT-Router: Effective and Interpretable Multi-LLM Routing via Item Response Theory

Wei Song, Zhenya Huang, Cheng Cheng, Weibo Gao, Bihan Xu, GuanHao Zhao, Fei Wang, Runze Wu

Large language models (LLMs) have demonstrated exceptional performance across a wide range of natural language tasks. However, selecting the optimal LLM to respond to a user query often necessitates a delicate balance between performance and cost. While powerful models deliver better results, they come at a high cost, whereas smaller models are more cost-effective but less capable. To address this trade-off, we propose IRT-Router, a multi-LLM routing framework that efficiently routes user queries to the most suitable LLM. Inspired by Item Response Theory (IRT), a psychological measurement methodology, IRT-Router explicitly models the relationship between LLM capabilities and user query attributes. This not only enables accurate prediction of response performance but also provides interpretable insights, such as LLM abilities and query difficulty. Additionally, we design an online query warm-up technique based on semantic similarity, further enhancing the online generalization capability of IRT-Router. Extensive experiments on 20 LLMs and 12 datasets demonstrate that IRT-Router outperforms most baseline methods in terms of effectiveness and interpretability. Its superior performance in cold-start scenarios further confirms the reliability and practicality of IRT-Router in real-world applications. Code is available at .

CodeDPO: Aligning Code Models with Self Generated and Verified Source Code

Kechi Zhang, Ge Li, Yihong Dong, Jingjing Xu, Jun Zhang, Jing Su, Yongfei Liu, Zhi Jin

Code generation models have shown significant potential for programming tasks. However, existing training methods like supervised fine-tuning face key limitations: they do not effectively teach models to prioritize correct over incorrect solutions in ambiguous situations, nor do they effectively optimize the runtime efficiency of the generated code. To address these challenges, we propose CodeDPO, a framework that integrates preference learning into code generation to improve two key code preference factors: code correctness and efficiency. CodeDPO employs a novel dataset construction method, utilizing a self-generation-and-validation mechanism that simultaneously generates and evaluates code and test cases. The underlying assumption is that test cases executable by multiple code snippets provide more reliable validation, and code that passes more tests is more likely to be correct. Through this self-validation process, our PageRank-inspired algorithm iteratively updates the ranking score of each code snippet, ultimately creating a code preference optimization dataset based on correctness and efficiency. CodeDPO is flexible and scalable, generating diverse preference optimization data without depending on powerful models such as GPT-4. Through comprehensive evaluations of five widely used benchmarks, CodeDPO demonstrates significant improvements in correctness and efficiency compared to existing methods. Our experiments prove that CodeDPO enhances the capabilities of LLMs in code generation and provides a robust foundation for conducting code preference optimization in more complex and challenging real-world scenarios.

Adaptive and Robust Translation from Natural Language to Multi-model Query Languages

Gengyuan Shi, Chaokun Wang, Liu Yabin, Jiawei Ren

Multi-model databases and polystore systems are increasingly studied for managing multi-model data holistically. As their primary interface, multi-model query languages (MMQLs) often exhibit complex grammars, highlighting the need for effective Text-to-MMQL translation methods. Despite advances in natural language translation, no effective solutions for Text-to-MMQL exist. To address this gap, we formally define the Text-to-MMQL task and present the first Text-to-MMQL dataset involving three representative MMQLs. We propose an adaptive Text-to-MMQL framework that includes both a schema embedding module for capturing multi-model schema information and an MMQL representation strategy to generate concise intermediate query formats with error correction in generated queries. Experimental results show that the proposed framework achieves over a 9% accuracy improvement over our adapted baseline methods.

RAEmoLLM: Retrieval Augmented LLMs for Cross-Domain Misinformation Detection Using In-Context Learning Based on Emotional Information

Zhiwei Liu, Kailai Yang, Qianqian Xie, Christine de Kock, Sophia Ananiadou, Eduard Hovy

Misinformation is prevalent in various fields such as education, politics, health, etc., causing significant harm to society. However, current methods for cross-domain misinformation detection rely on effort- and resource-intensive fine-tuning and complex model structures. With the outstanding performance of LLMs, many studies have employed them for misinformation detection. Unfortunately, they focus on in-domain tasks and do not incorporate significant sentiment and emotion features (which we jointly call {\em affect}). In this paper, we propose RAEmoLLM, the first retrieval augmented (RAG) LLMs framework to address cross-domain misinformation detection using in-context learning based on affective information. RAEmoLLM includes three modules. (1) In the index construction module, we apply an emotional LLM to obtain affective embeddings from all domains to construct a retrieval database. (2) The retrieval module uses the database to recommend top K examples (text-label pairs) from source domain data for target domain contents. (3) These examples are adopted as few-shot demonstrations for the inference module to process the target domain content. The RAEmoLLM can effectively enhance the general performance of LLMs in cross-domain misinformation detection tasks through affect-based retrieval, without fine-tuning. We evaluate our framework on three misinformation benchmarks. Results show that RAEmoLLM achieves significant improvements compared to the other few-shot methods on three datasets, with the highest increases of 15.64%, 31.18%, and 15.73% respectively. This project is available at https://github.com/lzw108/RAEmoLLM.

From Sub-Ability Diagnosis to Human-Aligned Generation: Bridging the Gap for Text Length Control via MarkerGen

Peiwen Yuan, Chuyi Tan, Shaoxiong Feng, Yiwei Li, Xinglin Wang, Yueqi Zhang, Jiayi Shi, Boyuan Pan, Yao Hu, Kan Li

Despite the rapid progress of large language models (LLMs), their length-controllable text generation (LCTG) ability remains below expectations, posing a major limitation for practical applications. Existing methods mainly focus on end-to-end training to reinforce adherence to length constraints. However, the lack of decomposition and targeted enhancement of LCTG sub-abilities restricts further progress. To bridge this gap, we conduct a bottom-up decomposition of LCTG sub-abilities with human patterns as reference and perform a detailed error analysis. On this basis, we propose MarkerGen, a simple-yet-effective plug-and-play approach that: (1) mitigates LLM fundamental deficiencies via external tool integration; (2) conducts explicit length modeling with dynamically inserted markers; (3) employs a three-stage generation scheme to better align length constraints while maintaining content quality. Comprehensive experiments demonstrate that MarkerGen significantly improves LCTG across various settings, exhibiting outstanding effectiveness and generalizability.

Hierarchical Attention Generates Better Proofs

Jianlong Chen, Chao Li, Yang Yuan, Andrew C Yao

Large language models (LLMs) have shown promise in formal theorem proving, but their token-level processing often fails to capture the inherent hierarchical nature of mathematical proofs. We introduce \textbf{Hierarchical Attention}, a regularization method that aligns LLMs' attention mechanisms with mathematical reasoning structures. Our approach establishes a five-level hierarchy from foundational elements to high-level concepts, ensuring structured information flow in proof generation. Experiments demonstrate that our method improves proof success rates by 2.05% on miniF2F and 1.69% on ProofNet while reducing proof complexity by 23.81% and 16.50% respectively. The code and models will be available.

Enhancing Text Editing for Grammatical Error Correction: Arabic as a Case Study

Bashar Alhafni, Nizar Habash

Text editing frames grammatical error correction (GEC) as a sequence tagging problem, where edit tags are assigned to input tokens, and applying these edits results in the corrected text. This approach has gained attention for its efficiency and interpretability. However, while extensively explored for English, text editing remains largely underexplored for morphologically rich languages like Arabic. In this paper, we introduce a text editing approach that derives edit tags directly from data, eliminating the need for language-specific edits. We demonstrate its effectiveness on Arabic, a diglossic and morphologically rich language, and investigate the impact of different edit representations on model performance. Our approach achieves SOTA results on two Arabic GEC benchmarks and performs on par with SOTA on two others. Additionally, our models are over six times faster than existing Arabic GEC systems, making our approach more practical for real-world applications. Finally, we explore ensemble models, demonstrating how combining different models leads to further performance improvements. We make our code, data, and pretrained models publicly available.

ELBA-Bench: An Efficient Learning Backdoor Attacks Benchmark for Large Language Models

Xuxu Liu, Siyuan Liang, Mengya Han, Yong Luo, Aishan Liu, Xiantao Cai, Zheng He, Dacheng Tao

Generative large language models are crucial in natural language processing, but they are vulnerable to backdoor attacks, where subtle triggers compromise their behavior. Although backdoor attacks against LLMs are constantly emerging, existing benchmarks remain limited in terms of sufficient coverage of attack, metric system integrity, backdoor attack alignment. And existing pre-trained backdoor attacks are idealized in practice due to resource access constraints. Therefore we establish \textit{ELBA-Bench}, a comprehensive and unified framework that allows attackers to inject backdoor through parameter efficient fine-tuning (\textit{e.g.,} LoRA) or without fine-tuning techniques (\textit{e.g.,} In-context-learning). \textit{ELBA-Bench} provides over 1300 experiments encompassing the implementations of 12 attack methods, 18 datasets, and 12 LLMs. Extensive experiments provide new invaluable findings into the strengths and limitations of various attack strategies. For instance, PEFT attack consistently outperform without fine-tuning approaches in classification tasks while showing strong cross-dataset generalization with optimized triggers boosting robustness; Task-relevant backdoor optimization techniques or attack prompts along with clean and adversarial demonstrations can enhance backdoor attack success while preserving model performance on clean samples. Additionally, we introduce a universal toolbox designed for standardized backdoor attack research at https://github.com/NWPUliuxx/ELBA_Bench, with the goal of propelling further progress in this vital area.

Mixture of Ordered Scoring Experts for Cross-prompt Essay Trait Scoring

Po-Kai Chen, Bo-Wei Tsai, Shao Kuan Wei, Chien-Yao Wang, Jia-Ching Wang, Yi-Ting Huang

Automated Essay Scoring (AES) plays a crucial role in language assessment. In particular, cross-prompt essay trait scoring provides learners with valuable feedback to improve their writing skills. However, due to the scarcity of prompts, most existing methods overlook critical information, such as content from prompts or essays, resulting in incomplete assessment perspectives. In this paper, we propose a robust AES framework, the Mixture of Ordered Scoring Experts (MOOSE), which integrates information from both prompts and essays. MOOSE employs three specialized experts to evaluate (1) the overall quality of an essay, (2) the relative quality across multiple essays, and (3) the relevance between an essay and its prompt. MOOSE introduces the ordered aggregation of assessment results from these experts along with effective feature learning techniques. Experimental results demonstrate that MOOSE achieves exceptionally stable and state-of-the-art performance in both cross-prompt scoring and multi-trait scoring on the ASAP++ dataset. The source code is released at https://github.com/antslabtw/MOOSE-AES.

ExploraCoder: Advancing Code Generation for Multiple Unseen APIs via Planning and Chained Exploration

Yunkun Wang, Yue Zhang, Zhen Qin, Chen Zhi, Binhua Li, Fei Huang, Yongbin Li, Shuiguang Deng

Large language models face intrinsic limitations in coding with APIs that are unseen in their training corpora. As libraries continuously evolve, it becomes impractical to exhaustively retrain LLMs with new API knowledge. This limitation hampers LLMs from solving programming problems which require newly introduced or privately maintained libraries. Inspired by exploratory programming paradigm in human behavior, we propose **ExploraCoder**, a training-free framework that empowers LLMs to invoke multiple unseen APIs in code solution by (1) planning a complex problem into several API invocation subtasks, and (2) experimenting with correct API usage at intermediate steps through a novel chain-of-API-exploration. We conduct evaluation on program synthesizing tasks involving complex API interactions. Experimental results demonstrate that ExploraCoder significantly improves performance for models lacking prior API knowledge, achieving absolute increases of up to 11.99\% over retrieval-based approaches and 17.28\% over pretraining-based methods in pass@10.

Can Indirect Prompt Injection Attacks Be Detected and Removed?

Yulin Chen, Haoran Li, Yuan Sui, Yufei He, Yue Liu, Yangqiu Song, Bryan Hooi

Prompt injection attacks manipulate large language models (LLMs) by misleading them to deviate from the original input instructions and execute maliciously injected instructions, because of their instruction-following capabilities and inability to distinguish between the original input instructions and maliciously injected instructions. To defend against such attacks, recent studies have developed various detection mechanisms. If we restrict ourselves specifically to works which perform detection rather than direct defense, most of them focus on direct prompt injection attacks, while there are few works for the indirect scenario, where injected instructions are indirectly from external tools, such as a search engine. Moreover, current works mainly investigate injection detection methods and pay less attention to the post-processing method that aims to mitigate the injection after detection. In this paper, we investigate the feasibility of detecting and removing indirect prompt injection attacks, and we construct a benchmark dataset for evaluation. For detection, we assess the performance of existing LLMs and open-source detection models, and we further train detection models using our crafted training datasets. For removal, we evaluate two intuitive methods: (1) the *segmentation removal method*, which segments the injected document and removes parts containing injected instructions, and (2) the *extraction removal method*, which trains an extraction model to identify and remove injected instructions.

Adversarial Alignment with Anchor Dragging Drift (A^3D^2): Multimodal Domain Adaptation with Partially Shifted Modalities

Jun Sun, Xinxin Zhang, Simin Hong, Jian Zhu, Lingfang Zeng

Multimodal learning has celebrated remarkable success across diverse areas, yet faces the challenge of prohibitively expensive data collection and annotation when adapting models to new environments. In this context, domain adaptation has gained growing popularity as a technique for knowledge transfer, which, however, remains underexplored in multimodal settings compared with unimodal ones. This paper investigates multimodal domain adaptation, focusing on a practical partially shifting scenario where some modalities (referred to as anchors) remain domain-stable, while others (referred to as drifts) undergo a domain shift. We propose a bi-alignment scheme to simultaneously perform drift-drift and anchor-drift matching. The former is achieved through adversarial learning, aligning the representations of the drifts across source and target domains; the latter corresponds to an "anchor dragging drift" strategy, which matches the distributions of the drifts and anchors within the target domain using the optimal transport (OT) method. The overall design principle features \textbf{A}dversarial \textbf{A}lignment with \textbf{A}nchor \textbf{D}ragging \textbf{D}rift, abbreviated as \textbf{A^3D^2}, for multimodal domain adaptation with partially shifted modalities. Comprehensive empirical results verify the effectiveness of the proposed approach, and demonstrate that A^3D^2 achieves superior performance compared with state-of-the-art approaches. The code is available at: .

DARS: Dynamic Action Re-Sampling to Enhance Coding Agent Performance by Adaptive Tree Traversal

Vaibhav Aggarwal, Ojasv Kamal, Abhinav Japesh, Zhijing Jin, Bernhard Schölkopf

Large Language Models (LLMs) have revolutionized various domains, including natural language processing, data analysis, and software development, by enabling automation. In software engineering, LLM-powered coding agents have garnered significant attention due to their potential to automate complex development tasks, assist in debugging, and enhance productivity. However, existing approaches often struggle with sub-optimal decision-making, requiring either extensive manual intervention or inefficient compute scaling strategies. To improve coding agent performance, we present Dynamic Action Re-Sampling (DARS), a novel inference time compute scaling approach for coding agents, that is faster and more effective at recovering from sub-optimal decisions compared to baselines. While traditional agents either follow linear trajectories or rely on random sampling for scaling compute, our approach DARS works by branching out a trajectory at certain key decision points by taking an alternative action given the history of the trajectory and execution feedback of the previous attempt from that point. We evaluate our approach on SWE-Bench Lite benchmark, demonstrating that this scaling strategy achieves a pass@k score of 55% with Claude 3.5 Sonnet V2. Our framework achieves a pass@1 rate of 47%, outperforming state-of-the-art (SOTA) open-source frameworks.

Targeted Syntactic Evaluation for Grammatical Error Correction

Aomi Koyama, Masato Mita, Su-Youn Yoon, Yasufumi Takama, Mamoru Komachi

Language learners encounter a wide range of grammar items across the beginner, intermediate, and advanced levels. To develop grammatical error correction (GEC) models effectively, it is crucial to identify which grammar items are easier or more challenging for models to correct. However, conventional benchmarks based on learner-produced texts are insufficient for conducting detailed evaluations of GEC model performance across a wide range of grammar items due to biases in their distribution. To address this issue, we propose a new evaluation paradigm that assesses GEC models using minimal pairs of ungrammatical and grammatical sentences for each grammar item. As the first benchmark within this paradigm, we introduce the CEFR-based Targeted Syntactic Evaluation Dataset for Grammatical Error Correction (CTSEG), which complements existing English benchmarks by enabling fine-grained analyses previously unattainable with conventional datasets. Using CTSEG, we evaluate three mainstream types of English GEC models: sequence-to-sequence models, sequence tagging models, and prompt-based models. The results indicate that while current models perform well on beginner-level grammar items, their performance deteriorates substantially for intermediate and advanced items.

Automated CAD Modeling Sequence Generation from Text Descriptions via Transformer-Based Large Language Models

JianXing Liao, Junyan Xu, Yatao Sun, Maowen Tang, Sicheng He, Jingxian Liao, Shui Yu, Yun Li, Xiaohong Guan

Designing complex computer-aided design (CAD) models is often time-consuming due to challenges such as computational inefficiency and the difficulty of generating precise models. We propose a novel language-guided framework for industrial design automation to address these issues, integrating large language models (LLMs) with computer-automated design (CAutoD).Through this framework, CAD models are automatically generated from parameters and appearance descriptions, supporting the automation of design tasks during the detailed CAD design phase. Our approach introduces three key innovations: (1) a semi-automated data annotation pipeline that leverages LLMs and vision-language large models (VLLMs) to generate high-quality parameters and appearance descriptions; (2) a Transformer-based CAD generator (TCADGen) that predicts modeling sequences via dual-channel feature aggregation; (3) an enhanced CAD modeling generation model, called CADLLM, that is designed to refine the generated sequences by incorporating the confidence scores from TCADGen. Experimental results demonstrate that the proposed approach outperforms traditional methods in both accuracy and efficiency, providing a powerful tool for automating industrial workflows and generating complex CAD models from textual prompts.The code is available at https://jianxliao.github.io/cadllm-page/

Can LLMs Reason About Program Semantics? A Comprehensive Evaluation of LLMs on Formal Specification Inference

Thanh Le-Cong, Bach Le, Toby Murray

Large Language Models (LLMs) are increasingly being used to automate programming tasks. However, the capabilities of LLMs in reasoning about program semantics are still inadequately studied, leaving substantial potential for further exploration. This paper introduces FormalBench, a comprehensive benchmark designed to evaluate the reasoning abilities of Large Language Models (LLMs) on program semantics. Specifically, it utilizes the task of synthesizing formal program specifications as a proxy measure for assessing the semantic reasoning of LLMs. This task requires both comprehensive reasoning over all possible program executions and the generation of precise, syntactically correct expressions that adhere to formal syntax and semantics. Using this benchmark, we evaluated the ability of LLMs to synthesize consistent and complete specifications. Our findings show that LLMs perform well with simple control flows but struggle with more complex structures, especially loops, even with advanced prompting. Additionally, LLMs exhibit limited robustness against semantic-preserving transformations. We also highlight common failure patterns and design self-repair prompts, improving success rates by 25%. FormalBench is packaged as an executable library and has been released at https://github.com/thanhlecongg/FormalBench/.

Decoder-Only LLMs can be Masked Auto-Encoders

Dan Qiao, Yuan Gao, Zheming Yang, Di Yang, Ziheng Wu, Pengcheng Lu, Minghui Qiu, Juntao Li, Min Zhang

Modern NLP workflows (e.g., RAG systems) require different models for generation and embedding tasks, where bidirectional pre-trained encoders and decoder-only Large Language Models (LLMs) dominate respective tasks. Structural differences between models result in extra development costs and limit knowledge sharing between tasks. In this work, we present UniMAE, a novel unsupervised training method that transforms an Decoder-Only LLM into a \textbf{Uni}-Directional \textbf{M}asked \textbf{A}uto-\textbf{E}ncoder. UniMAE compresses high-quality semantic information into the [EOS] embedding while preserving the generation capabilities of LLMs. Comprehensive evaluations across 56 MTEB datasets demonstrate that UniMAE can achieve state-of-the-art results under unsupervised settings with merely 100 training steps, establishing the first effective approach to unifying generation and representation learning in decoder-only architectures.

Boosting LLM’s Molecular Structure Elucidation with Knowledge Enhanced Tree Search Reasoning

Xiang Zhuang, Bin Wu, Jiyu Cui, Kehua Feng, Xiaotong Li, Huabin Xing, Keyan Ding, Qiang Zhang, Huajun Chen

Molecular structure elucidation involves deducing a molecule's structure from various types of spectral data, which is crucial in chemical experimental analysis. While large language models (LLMs) have shown remarkable proficiency in analyzing and reasoning through complex tasks, they still encounter substantial challenges in molecular structure elucidation. We identify that these challenges largely stem from LLMs' limited grasp of specialized chemical knowledge. In this work, we introduce a Knowledge-enhanced reasoning framework for Molecular Structure Elucidation (K-MSE), leveraging Monte Carlo Tree Search for test-time scaling as a plugin. Specifically, we construct an external molecular substructure knowledge base to extend the LLMs' coverage of the chemical structure space. Furthermore, we design a specialized molecule-spectrum scorer to act as a reward model for the reasoning process, addressing the issue of inaccurate solution evaluation in LLMs. Experimental results show that our approach significantly boosts performance, particularly gaining more than 20% improvement on both GPT-4o-mini and GPT-4o.

ChemActor: Enhancing Automated Extraction of Chemical Synthesis Actions with LLM-Generated Data

Yu Zhang, Ruijie Yu, Jidong Tian, Feng Zhu, Jiapeng Liu, Xiaokang Yang, Yaohui Jin, Yanyan Xu

With the increasing interest in robotic synthesis in the context of organic chemistry, the automated extraction of chemical procedures from literature is critical. However, this task remains challenging due to the inherent ambiguity of chemical language and the high cost of human annotation required for developing reliable computer-aided extraction protocols. Here, we present \textbf{ChemActor}, a fully fine-tuned large language model (LLM), as a chemical executor to convert between unstructured experimental procedures and structured action sequences. We propose a sequential LLM-generated data framework to address the challenges of insufficient and low-quality annotated data. This framework integrates a data selection module that selects data based on distribution divergence, with a general-purpose LLM, to generate machine-executable actions from a single molecule input. Additionally, we introduce a novel multi-round LLMs circle review metric, which reflects the model's advanced understanding of chemical experimental procedures. Extensive experiments on reaction-to-description (R2D) and description-to-action (D2A) tasks demonstrate that ChemActor, augmented by LLM-generated data, achieves state-of-the-art performance, outperforming the baseline model by 10\%. The code is available at: https://github.com/Zhanghahah/ChemActor.

ALGEN: Few-shot Inversion Attacks on Textual Embeddings using Alignment and Generation

Yiyi Chen, Qiongkai Xu, Johannes Bjerva

With the growing popularity of Large Language Models (LLMs) and vector databases, private textual data is increasingly processed and stored as numerical embeddings. However, recent studies have proven that such embeddings are vulnerable to inversion attacks, where original text is reconstructed to reveal sensitive information. Previous research has largely assumed access to millions of sentences to train attack models, e.g., through data leakage or nearly unrestricted API access. With our method, a single data point is sufficient for a partially successful inversion attack. With as little as 1k data samples, performance reaches an optimum across a range of black-box encoders, without training on leaked data. We present a Few-shot Textual Embedding Inversion Attack using Cross-Model **AL**ignment and **GEN**eration (__ALGEN__), by aligning victim embeddings to the attack space and using a generative model to reconstruct text. We find that __ALGEN__ attacks can be effectively transferred across domains and languages, revealing key information. We further examine a variety of defense mechanisms against **ALGEN**, and find that none are effective, highlighting the vulnerabilities posed by inversion attacks. By significantly lowering the cost of inversion and proving that embedding spaces can be aligned through one-step optimization, we establish a new textual embedding inversion paradigm with broader applications for embedding alignment in NLP.

Delta-KNN: Improving Demonstration Selection in In-Context Learning for Alzheimer's Disease Detection

Chuyuan Li, Raymond Li, Thalia S. Field, Giuseppe Carenini

Alzheimer’s Disease (AD) is a progressive neurodegenerative disorder that leads to dementia, and early intervention can greatly benefit from analyzing linguistic abnormalities. In this work, we explore the potential of Large Language Models as health assistants for AD diagnosis from patient-generated text using in-context learning (ICL), where tasks are defined through a few input-output examples. Empirical results reveal that conventional ICL methods, such as similarity-based selection, perform poorly for AD diagnosis, likely due to the inherent complexity of this task. To address this, we introduce Delta-KNN, a novel demonstration selection strategy that enhances ICL performance. Our method leverages a delta score to assess the relative gains of each training example, coupled with a KNN-based retriever that dynamically selects optimal ``representatives'' for a given input. Experiments on two AD detection datasets across three models demonstrate that Delta-KNN consistently outperforms existing ICL baselines. Notably, when using the Llama-3.1 model, our approach achieves new state-of-the-art results, surpassing even supervised classifiers.

From Informal to Formal – Incorporating and Evaluating LLMs on Natural Language Requirements to Verifiable Formal Proofs

Jialun Cao, Yaojie Lu, Meiziniu Li, Haoyang Ma, Haokun Li, Mengda He, Cheng Wen, Le Sun, Hongyu Zhang, Shengchao Qin, Shing-Chi Cheung, Cong Tian

The research in AI-based formal mathematical reasoning has shown an unstoppable growth trend. These studies have excelled in mathematical competitions like IMO and have made significant progress. However, these studies intertwined multiple skills simultaneously—problem-solving, reasoning, and writing formal specifications—making it hard to precisely identify the LLMs' strengths and weaknesses in each task. This paper focuses on formal verification, an immediate application scenario of formal reasoning, and breaks it down into sub-tasks. We constructed 18k high-quality instruction-response pairs across five mainstream formal specification languages (Coq, Lean4, Dafny, ACSL, and TLA+) in six tasks by distilling gpt-4o and evaluated against ten open-sourced LLMs, including recent popular DeepSeek-R1. We found that LLMs are good at writing proof segments when given either the code, or the detailed description of proof steps. Also, the fine-tuning brought about a nearly threefold improvement at most. And interestingly, we observed that fine-tuning with formal data also enhances abilities in mathematics, reasoning, and coding. We hope our findings inspire further research.

CheXalign: Preference fine-tuning in chest X-ray interpretation models without human feedback

Dennis Hein, Zhihong Chen, Sophie Ostmeier, Justin Xu, Maya Varma, Eduardo Pontes Reis, Arne Edward Michalson MD, Christian Bluethgen, Hyun Joo Shin, Curtis Langlotz, Akshay S Chaudhari

Radiologists play a crucial role in translating medical images into actionable reports. However, the field faces staffing shortages and increasing workloads. While automated approaches using vision-language models (VLMs) show promise as assistants, they require exceptionally high accuracy. Most current VLMs in radiology rely solely on supervised fine-tuning. Meanwhile, additional preference fine-tuning in the post-training pipeline has become standard practice in the general domain. The challenge in radiology lies in the prohibitive cost of obtaining radiologist feedback at scale. To address this challenge, we propose an automated pipeline for preference feedback, focusing on chest X-ray radiology report generation (RRG). Specifically, our method leverages publicly available datasets containing pairs of images and radiologist-written reference reports with reference-based metrics, or Judges, eliminating the need for *additional radiologist feedback*. We investigate reward overoptimization via length exploitation in this setting and introduce a length-controlled version of the GREEN score. Our best-performing setup achieves state-of-the-art CheXbert scores on the MIMIC-CXR dataset for the RRG task while on average maintaining robust performance across six additional image perception and reasoning tasks.

MIR: Methodology Inspiration Retrieval for Scientific Research Problems

Aniketh Garikaparthi, Manasi Patwardhan, Aditya Sanjiv Kanade, Aman Hassan, Lovekesh Vig, Arman Cohan

There has been a surge of interest in harnessing the reasoning capabilities of Large Language Models (LLMs) to accelerate scientific discovery. While existing approaches rely on grounding the discovery process within the relevant literature, effectiveness varies significantly with the quality and nature of the retrieved literature. We address the challenge of retrieving prior work whose concepts can inspire solutions for a given research problem, a task we define as Methodology Inspiration Retrieval (MIR). We construct a novel dataset tailored for training and evaluating retrievers on MIR, and establish baselines. To address MIR, we build the Methodology Adjacency Graph (MAG); capturing methodological lineage through citation relationships. We leverage MAG to embed an “intuitive prior’’ into dense retrievers for identifying patterns of methodological inspiration beyond superficial semantic similarity. This achieves significant gains of +5.4 in Recall@3 and +7.8 in Mean Average Precision (mAP) over strong baselines. Further, we adapt LLM-based re-ranking strategies to MIR, yielding additional improvements of +4.5 in Recall@3 and +4.8 in mAP. Through extensive ablation studies and qualitative analyses, we exhibit the promise of MIR in enhancing automated scientific discovery and outline avenues for advancing inspiration-driven retrieval.

Robust Utility-Preserving Text Anonymization Based on Large Language Models

Tianyu Yang, Xiaodan Zhu, Iryna Gurevych

Anonymizing text that contains sensitive information is crucial for a wide range of applications. Existing techniques face the emerging challenges of the re-identification ability of large language models (LLMs), which have shown advanced capability in memorizing detailed information and reasoning over dispersed pieces of patterns to draw conclusions. When defending against LLM-based re-identification, anonymization could jeopardize the utility of the resulting anonymized data in downstream tasks. In general, the interaction between anonymization and data utility requires a deeper understanding within the context of LLMs. In this paper, we propose a framework composed of three key LLM-based components: \textit{a privacy evaluator}, \textit{a utility evaluator} and \textit{an optimization component}, which work collaboratively to perform anonymization. Extensive experiments demonstrate that the proposed model outperforms existing baselines, showing robustness in reducing the risk of re-identification while preserving greater data utility in downstream tasks. We provide detailed studies on these core modules. To consider large-scale and real-time applications, we investigate the distillation of the anonymization capabilities into lightweight models. All of our code and datasets will be made publicly available at \texttt{[Github URL]}.

CoachMe: Decoding Sport Elements with a Reference-Based Coaching Instruction Generation Model

Wei-Hsin Yeh, Yu-An Su, Chih-Ning Chen, Yi-Hsueh Lin, Calvin Ku, WENHSIN CHIU, Min-Chun Hu, Lun-Wei Ku

Motion instruction is a crucial task that helps athletes refine their technique by analyzing movements and providing corrective guidance. Although recent advances in multimodal models have improved motion understanding, generating precise and sport-specific instruction remains challenging due to the highly domain-specific nature of sports and the need for informative guidance. We propose CoachMe, a reference-based model that analyzes the differences between a learner's motion and a reference under temporal and physical aspects. This approach enables both domain-knowledge learning and the acquisition of a coach-like thinking process that identifies movement errors effectively and provides feedback to explain how to improve. In this paper, we illustrate how CoachMe adapts well to specific sports such as skating and boxing by learning from general movements and then leveraging limited data. Experiments show that CoachMe provides high-quality instructions instead of directions merely in the tone of a coach but without critical information. CoachMe outperforms GPT-4o by 31.6\% in G-Eval on figure skating and by 58.3\% on boxing. Analysis further confirms that it elaborates on errors and their corresponding improvement methods in the generated instructions. You can find CoachMe here:

Hierarchical Memory Organization for Wikipedia Generation

Eugene J. Yu, Dawei Zhu, Yifan Song, Xiangyu Wong, Jiebin Zhang, Wenxuan Shi, Xiaoguang Li, Qun Liu, Sujian Li

Generating Wikipedia articles autonomously is a challenging task requiring the integration of accurate, comprehensive, and well-structured information from diverse sources. This paper introduces the Memory Organization-based Generation (MOG) framework, a novel approach to address these challenges by leveraging a hierarchical memory architecture. MOG extracts fine-grained memory units from web documents, recursively organizes them into a Wikipedia-style hierarchical structure, and uses this structure to guide the generation process. This ensures alignment between memory and the article outline, improving both informativeness and verifiability while minimizing hallucinations. Additionally, a citation module is implemented to enhance traceability by linking every generated sentence to specific memory units. Evaluations on our newly created WikiStart dataset demonstrate that MOG outperforms baseline methods in producing informative and reliable articles, making it particularly robust in real-world scenarios.

Time-MQA: Time Series Multi-Task Question Answering with Context Enhancement

Yaxuan Kong, Yiyuan Yang, Yoontae Hwang, Wenjie Du, Stefan Zohren, Zhangyang Wang, Ming Jin, Qingsong Wen

Time series data are foundational in finance, healthcare, and energy domains. However, most existing methods and datasets remain focused on a narrow spectrum of tasks, such as forecasting or anomaly detection. To bridge this gap, we introduce Time Series Multi-Task Question Answering (Time-MQA), a unified framework that enables natural language queries across multiple time series tasks - numerical analytical tasks and open-ended question answering with reasoning. Central to Time-MQA is the TSQA dataset, a large-scale dataset containing \sim200k question-answer pairs derived from diverse time series spanning environment, traffic, etc. This comprehensive resource covers various time series lengths and promotes robust model development. We further demonstrate how continually pre-training large language models (Mistral 7B, Llama-3 8B, and Qwen-2.5 7B) on the TSQA dataset enhanced time series reasoning capabilities, moving beyond mere numeric tasks and enabling more advanced and intuitive interactions with temporal data. The complete TSQA dataset, models, user study questionnaires for evaluation, and other related materials have been open-sourced here.

ETF: An Entity Tracing Framework for Hallucination Detection in Code Summaries

Kishan Maharaj, Vitobha Munigala, Srikanth G. Tamilselvam, Prince Kumar, Sayandeep Sen, Palani Kodeswaran, Abhijit Mishra, Pushpak Bhattacharyya

Recent advancements in large language models (LLMs) have significantly enhanced their ability to understand both natural language and code, driving their use in tasks like natural language-to-code (NL2Code) and code summarisation. However, LLMs are prone to hallucination—outputs that stray from intended meanings. Detecting hallucinations in code summarisation is especially difficult due to the complex interplay between programming and natural languages. We introduce a first-of-its-kind dataset, CodeSumEval, with \sim10K samples, curated specifically for hallucination detection in code summarisation. We further propose a novel Entity Tracing Framework (ETF) that a) utilises static program analysis to identify code entities from the program and b) uses LLMs to map and verify these entities and their intents within generated code summaries. Our experimental analysis demonstrates the framework's effectiveness, leading to a 73\% F1 score. The proposed approach provides a method for detecting hallucinations by tracing entities from the summary to the code, allowing us to evaluate summary accuracy and localise the error within the summary.

Representation-based Reward Modeling for Efficient Safety Alignment of Large Language Model

Deng Qiyuan, Xuefeng Bai, Kehai Chen, Yaowei Wang, Liqiang Nie, Min Zhang

Reinforcement Learning (RL) algorithms for safety alignment of Large Language Models (LLMs), such as Direct Preference Optimization (DPO), encounter the challenge of distribution shift. Current approaches typically address this issue through online sampling from the target policy, which requires significant computational resources. In this paper, we hypothesize that during off-policy training, while the ranking order of output generated by policy changes, their overall distribution remains relatively stable. This stability allows the conversion of the sampling process from the target policy into a computationally efficient re-ranking of preference data. Building on this hypothesis, we propose a new framework that leverages the model's intrinsic safety judgment capability to extract reward signals, which are then used to calculate label confidence for preference reordering. Extensive experiments and theoretical analysis demonstrate that the proposed method effectively addresses the distribution shift issue, remarkably enhancing the safety performance while avoiding about 300x computational overheads.

HiAgent: Hierarchical Working Memory Management for Solving Long-Horizon Agentic Tasks with Large Language Models

Mengkang Hu, Tianxing Chen, Qiguang Chen, Yao Mu, Wenqi Shao, Ping Luo

Large Language Model (LLM)-based agents exhibit significant potential across various domains, operating as interactive systems that process environmental observations to generate executable actions for target tasks. The effectiveness of these agents is significantly influenced by their memory mechanism, which records historical experiences as sequences of action-observation pairs. We categorize memory into two types: cross-trial memory, accumulated across multiple attempts, and in-trial memory (working memory), accumulated within a single attempt. While considerable research has optimized performance through cross-trial memory, the enhancement of agent performance through improved working memory utilization remains underexplored. Instead, existing approaches often involve directly inputting entire historical action-observation pairs into LLMs, leading to redundancy in long-horizon tasks. Inspired by human problem-solving strategies, this paper introduces HiAgent, a framework that leverages subgoals as memory chunks to manage the working memory of LLM-based agents hierarchically. Specifically, HiAgent prompts LLMs to formulate subgoals before generating executable actions and enables LLMs to decide proactively to replace previous subgoals with summarized observations, retaining only the action-observation pairs relevant to the current subgoal. Experimental results across five long-horizon tasks demonstrate that HiAgent achieves a twofold increase in success rate and reduces the average number of steps required by 3.8. Additionally, our analysis shows that HiAgent consistently improves performance across various steps, highlighting its robustness and generalizability. Code is available in this URL: https://github.com/HiAgent2024/HiAgent

Session 7: Question Answering (Tuesday 29 July, 10:30-12:00, Hall 4/5)

GRAF: Graph Retrieval Augmented by Facts for Romanian Legal Multi-Choice Question Answering

Cristian-George Craciun, Răzvan-Alexandru Smădu, Dumitru-Clementin Cercel, Mihaela-Claudia Cercel

Pre-trained language models have shown remarkable performances in recent years, setting a new paradigm for natural language processing (NLP) research. The legal domain has received some attention from the NLP community partly due to its textual nature. Question-answering (QA) systems represent some of the tasks in this domain. This work explores the legal Multiple-Choice QA (MCQA) for a low-resource language, Romanian. The contribution of this work is multi-fold. We introduce JuRO, the first openly available Romanian legal MCQA dataset, comprising 10,836 questions from three examinations. Along with this dataset, we introduce CROL, an organized corpus of laws with a total of 93 distinct documents with their modifications over 763 time spans, which we used for information retrieval techniques in this work. Additionally, we construct Law-RoG, the first graph of legal knowledge for the Romanian language derived from the corpus mentioned above. Lastly, we propose a novel approach for MCQA, namely Graph Retrieval Augmented by Facts (GRAF), which achieves competitive results with generally accepted state-of-the-art methods and even exceeds them in most settings.

PECAN: LLM-Guided Dynamic Progress Control with Attention-Guided Hierarchical Weighted Graph for Long-Document QA

Xinyu Wang, Yanzheng Xiang, Lin Gui, Yulan He

Long-document QA presents challenges with large-scale text and long-distance dependencies. Recent advances in Large Language Models (LLMs) enable entire documents to be processed in a single pass. However, their computational cost is significantly high. Retrieval-Augmented Generation (RAG) methods split text into smaller chunks, but they often yield inferior results and may lose global context. Recent approaches that integrate LLMs into RAG via iterative summarization either underutilize LLM capabilities or still incur high computational costs. In this paper, we combine the high accuracy of LLMs with the efficiency of RAG and propose LLM-Guided Dynamic Progress Control with Attention-Based Hierarchical Weighted Graph (PECAN). Our method introduces two key improvements: (1) LLM-Guided Dynamic Progress Control: We leverage LLMs to dynamically control the retrieval process, adjusting the amount of retrieved information based on different queries to achieve a better balance of effectiveness and efficiency. (2) Attention-Guided Retrieval: We propose a novel retrieval method that constructs a hierarchical graph where edges are derived by LLM attention weights. Experimental results demonstrate that PECAN achieves LLM-level performance while maintaining computational complexity comparable to that of RAG methods on two single-document and two multi-document QA datasets.

Lightweight Query Checkpoint: Classifying Faulty User Queries to Mitigate Hallucinations in Large Language Model Question Answering

Jonghak Jang, Minjoo Son, Misuk Kim

Question Answering (QA) with large language models has shown impressive performance, yet hallucinations still persist, particularly when user queries carry incorrect premises, insufficient context, or linguistic ambiguity. To address this issue, we propose Lightweight Query Checkpoint (LQC), a small classification model that detects verification-required queries before the LLM generates a potentially faulty answer. LQC leverages hidden states extracted from intermediate layers of a smaller-scale, non-instruct-tuned LLM to effectively distinguish queries requiring verification from clear queries. We first systematically define categories of queries that need verification, construct a dataset comprising both defective and clear queries, and train a binary contrastive learning model. Through extensive experiments on various QA datasets, we demonstrate that incorporating LQC into QA pipelines reduces hallucinations while preserving strong answer quality.

A Study into Investigating Temporal Robustness of LLMs

Jonas Wallat, Abdelrahman Abdallah, Adam Jatowt, Avishek Anand

Large Language Models (LLMs) encapsulate a surprising amount of factual world knowledge. However, their performance on temporal questions and historical knowledge is limited because they often cannot understand temporal scope and orientation or neglect the temporal aspect altogether. In this study, we aim to measure precisely how robust LLMs are for question answering based on their ability to process temporal information and perform tasks requiring temporal reasoning and temporal factual knowledge. Specifically, we design eight time-sensitive robustness tests for factual information to check the sensitivity of six popular LLMs in the zero-shot setting. Overall, we find LLMs lacking temporal robustness, especially to temporal reformulations and the use of different granularities of temporal references. We show how a selection of these eight tests can be used automatically to judge a model's temporal robustness for user questions on the fly. Finally, we apply the findings of this study to improve the temporal QA performance by up to 55\%.

ASTRID - An Automated and Scalable TRIaD for the Evaluation of RAG-based Clinical Question Answering Systems

Yajie Vera He, Mohita Chowdhury, Jared Joselowitz, Aisling Higham, Ernest Lim

Large Language Models (LLMs) have shown impressive potential in clinical question answering (QA), with Retrieval Augmented Generation (RAG) emerging as a leading approach for ensuring the factual accuracy of model responses. However, current automated RAG metrics perform poorly in clinical and conversational use cases. Using clinical human evaluations of responses is expensive, unscalable, and not conducive to the continuous iterative development of RAG systems. To address these challenges, we introduce ASTRID - an Automated and Scalable TRIaD for evaluating clinical QA systems leveraging RAG - consisting of three metrics: Context Relevance (CR), Refusal Accuracy (RA), and Conversational Faithfulness (CF). Our novel evaluation metric, CF, is designed to better capture the faithfulness of a model's response to the knowledge base without penalising conversational elements. Additionally, our metric RA captures the refusal to address questions outside of the system's scope of practice. To validate our triad, we curate a dataset of over 200 real-world patient questions posed to an LLM-based QA agent during surgical follow-up for cataract surgery - the highest volume operation in the world - augmented with clinician-selected questions for emergency, and clinical and non-clinical out-of-domain scenarios. We demonstrate that CF predicts human ratings of faithfulness more accurately than existing definitions in conversational settings. Furthermore, using eight different LLMs, we demonstrate that the three metrics can closely agree with human evaluations, highlighting the potential of these metrics for use in LLM-driven automated evaluation pipelines. Finally, we show that evaluation using our triad of CF, RA, and CR exhibits alignment with clinician assessment for inappropriate, harmful, or unhelpful responses. We also publish the prompts and datasets for these experiments, providing valuable resources for further research and development.

Disambiguate First Parse Later: Generating Interpretations for Ambiguity Resolution in Semantic Parsing

Irina Saparina, Mirella Lapata

Handling ambiguity and underspecification is an important challenge in natural language interfaces, particularly for tasks like text-to-SQL semantic parsing. We propose a modular approach that resolves ambiguity using natural language interpretations before mapping these to logical forms (e.g., SQL queries). Although LLMs excel at parsing unambiguous utterances, they show strong biases for ambiguous ones, typically predicting only preferred interpretations. We constructively exploit this bias to generate an initial set of preferred disambiguations and then apply a specialized infilling model to identify and generate missing interpretations. To train the infilling model, we introduce an annotation method that uses SQL execution to validate different meanings. Our approach improves interpretation coverage and generalizes across datasets with different annotation styles, database structures, and ambiguity types.

Generate, Discriminate, Evolve: Enhancing Context Faithfulness via Fine-Grained Sentence-Level Self-Evolution

Kun LI, Tianhua Zhang, Yunxiang Li, Hongyin Luo, Abdalla Mohamed Salama Sayed Moustafa, Xixin Wu, James R. Glass, Helen M. Meng

Improving context faithfulness in large language models is essential for developing trustworthy retrieval augmented generation systems and mitigating hallucinations, especially in long-form question answering (LFQA) tasks or scenarios involving knowledge conflicts. Existing methods either intervene LLMs only at inference without addressing their inherent limitations or overlook the potential for self-improvement. In this paper, we introduce GenDiE(Generate, Discriminate, Evolve), a novel self-evolving framework that enhances context faithfulness through fine-grained sentence-level optimization. GenDiE combines both generative and discriminative training, equipping LLMs with self-generation and self-scoring capabilities to facilitate iterative self-evolution. This supports both data construction for model alignment and score-guided search during inference. Furthermore, by treating each sentence in a response as an independent optimization unit, GenDiE effectively addresses the limitations of previous approaches that optimize at the holistic answer level, which may miss unfaithful details. Experiments on ASQA (in-domain LFQA) and ConFiQA (out-of-domain counterfactual QA) datasets demonstrate that GenDiE surpasses various baselines in both faithfulness and correctness, and exhibits robust performance for domain adaptation.

Enhancing Multi-Hop Reasoning for Question Answering with Hyperbolic Representations

Simon Welz, Lucie Flek, Akbar Karimi

Hyperbolic representations are effective in modeling knowledge graph data which is prevalently used to facilitate multi-hop reasoning. However, a rigorous and detailed comparison of the two spaces for this task is lacking. In this paper, through a simple integration of hyperbolic representations with an encoder-decoder model, we perform a controlled and comprehensive set of experiments to compare the capacity of hyperbolic space versus Euclidean space in multi-hop reasoning. Our results show that the former consistently outperforms the latter across a diverse set of datasets. In addition, through an ablation study, we show that a learnable curvature initialized with the delta hyperbolicity of the utilized data yields superior results to random initializations. Furthermore, our findings suggest a direct correlation between the dataset difficulty, defined as the number of out-going relations in a knowledge graph, and the effectiveness of the hyperbolic space, indicating that hyperbolic representations are more beneficial for harder datasets.

Recursive Question Understanding for Complex Question Answering over Heterogeneous Personal Data

Philipp Christmann, Gerhard Weikum

Question answering over mixed sources, like text and tables, has been advanced by verbalizing all contents and encoding it with a language model. A prominent case of such heterogeneous data is personal information: user devices log vast amounts of data every day, such as calendar entries, workout statistics, shopping records, streaming history, and more. Information needs range from simple look-ups to queries of analytical nature. The challenge is to provide humans with convenient access with small footprint, so that all personal data stays on the user devices. We present ReQAP, a novel method that creates an executable operator tree for a given question, via recursive decomposition. Operators are designed to enable seamless integration of structured and unstructured sources, and the execution of the operator tree yields a traceable answer. We further release the PerQA benchmark, with persona-based data and questions, covering a diverse spectrum of realistic user needs.

LLM-Symbolic Integration for Robust Temporal Tabular Reasoning

Atharv Kulkarni, Kushagra Dixit, Vivek Srikumar, Dan Roth, Vivek Gupta

Temporal tabular question answering presents a significant challenge for Large Language Models (LLMs), requiring robust reasoning over structured data—a task where traditional prompting methods often fall short. These methods face challenges such as memorization, sensitivity to table size, and reduced performance on complex queries. To overcome these limitations, we introduce TEMPTABQA-C, a synthetic dataset designed for systematic and controlled evaluations, alongside a symbolic intermediate representation that transforms tables into database schemas. This structured approach allows LLMs to generate and execute SQL queries, enhancing generalization and mitigating biases. By incorporating adaptive fewshot prompting with contextually tailored examples, our method achieves superior robustness, scalability, and performance. Experimental results consistently highlight improvements across key challenges, setting a new benchmark for robust temporal reasoning with LLMs.

Verbosity-Aware Rationale Reduction: Sentence-Level Rationale Reduction for Efficient and Effective Reasoning

Joonwon Jang, Jaehee Kim, WONBIN KWEON, Seonghyeon Lee, Hwanjo Yu

Large Language Models (LLMs) rely on generating extensive intermediate reasoning units (e.g., tokens, sentences) to enhance final answer quality across a wide range of complex tasks. While this approach has proven effective, it inevitably increases substantial inference costs. Previous methods adopting token-level reduction without clear criteria result in poor performance compared to models trained with complete rationale. To address this challenge, we propose a novel sentence-level rationale reduction framework leveraging likelihood-based criteria, *verbosity*, to identify and remove redundant reasoning sentences. Unlike previous approaches, our method leverages *verbosity* to selectively remove redundant reasoning sentences while preserving reasoning capabilities. Our experimental results across various reasoning tasks demonstrate that our method improves performance by an average of 7.71% while reducing token generation by 19.87% compared to model trained with complete reasoning paths.

Learning to Insert [PAUSE] Tokens for Better Reasoning

Eunki Kim, Sangryul Kim, James Thorne

To enhance reasoning capabilities, research has explored incorporating special-purpose tokens into the training process to strengthen the learning mechanism of transformer-based LLMs. Building on prior research that showed inserting dummy tokens consecutively just before reasoning steps can enhance effectiveness, we introduce a novel approach termed Dynamic Inserting tokens Training (DIT). Our method identifies positions within sequences where model confidence is lowest identified by token log likelihood, strategically inserting [PAUSE] tokens to bolster the model's predictive capabilities for subsequent tokens. Experimental results across diverse datasets and models, from 2.7B model to 8B model, demonstrate that DIT consistently outperforms traditional fine-tuning and previous token insertion methods. With this simple yet impactful method, we achieve accuracy gains of up to 4.7%p on GSM8K, 3.23%p on AQUA-RAT, and pass@1 improvements of up to 3.4%p on MBPP. Our work shows a model-based, dynamic approach rather than a heuristic one, thereby broadening the scope of research in reasoning.

MDBench: A Synthetic Multi-Document Reasoning Benchmark Generated with Knowledge Guidance

Joseph J Peper, Wenzhao Qiu, Ali Payani, Lu Wang

Natural language processing evaluation has made significant progress, largely driven by the proliferation of powerful large language mod-els (LLMs). New evaluation benchmarks are of increasing priority as the reasoning capabilities of LLMs are expanding at a rapid pace. In particular, while multi-document (MD) reasoning is an area of extreme relevance given LLM capabilities in handling longer-context inputs, few benchmarks exist to rigorously examine model behavior in this setting. Moreover, the multi-document setting is historically challenging for benchmark creation due to the expensive cost of annotating long inputs. In this work, we introduce MDBench, a new dataset for evaluating LLMs on the task of multi-document reasoning. Notably, MDBench is created through a novel synthetic generation process, allowing us to controllably and efficiently generate challenging document sets and the corresponding question-answer (QA) examples. Our novel technique operates on condensed structured seed knowledge, modifying it through LLM-assisted edits to induce MD-specific reasoning challenges. We then convert this structured knowledge into a natural text surface form, generating a document set and corresponding QA example. We analyze the behavior of popular LLMs and prompting techniques, finding that MDBench poses significant challenges for all methods, even with relatively short document sets. We also see our knowledge-guided generation technique (1) allows us to readily perform targeted analysis of MD-specific reasoning capabilities and (2) can be adapted quickly to account for new challenges and future modeling improvements.

Search-in-Context: Efficient Multi-Hop QA over Long Contexts via Monte Carlo Tree Search with Dynamic KV Retrieval

Jiabei Chen, Guang Liu, Shizhu He, Kun Luo, Yao Xu, Jun Zhao, Kang Liu

Recent advancements in large language models (LLMs) have demonstrated remarkable capabilities in complex reasoning tasks, such as math problem-solving and code generation. However, multi-hop question answering (MHQA) over long contexts, which demands both robust knowledge-intensive reasoning and efficient processing of lengthy documents, remains a significant challenge. Existing approaches often struggle to balance these requirements, either neglecting explicit reasoning or incurring expensive computational costs due to full-attention mechanisms over long contexts. To address this, we propose **Search-in-Context (SIC)**, a novel framework that integrates Monte Carlo Tree Search (MCTS) with dynamic key-value (KV) retrieval to enable iterative, context-aware reasoning. SIC dynamically retrieves critical KV pairs (e.g., 4K tokens) at each step, prioritizing relevant evidence while mitigating the "lost in the middle" problem. Furthermore, the paper introduces a Process-Reward Model (PRM) trained on auto-labeled data to guide the MCTS process with stepwise rewards, promoting high-quality reasoning trajectories without manual annotation. Experiments on three long-context MHQA benchmarks (HotpotQA, 2WikiMultihopQA, MuSiQue) and a counterfactual multi-hop dataset demonstrate SIC's superiority, achieving state-of-the-art performance while significantly reducing computational overhead.

Enhancing Complex Reasoning in Knowledge Graph Question Answering Through Query Graph Approximation

Hongjun Jeong, Minji Kim, Heesoo Jung, Ko Keun Kim, Hogun Park

Knowledge-grounded question answering (QA) leverages knowledge graphs (KGs) to provide answers based on structured factual data. Existing approaches are mainly divided into Knowledge Graph Question Answering (KGQA) and Complex Query Answering (CQA). Both approaches have limitations: the first struggles when essential triplets related to the questions are missing in the given KG, while the second depends on structured first-order logic queries. To overcome these limitations, we propose a novel framework, termed Aqua-QA. Aqua-QA approximates query graphs from natural language questions, enabling reasoning over KGs. We evaluate Aqua-QA on challenging QA tasks where the entities or relations representing the input questions are excluded from the KGs, and complex logical reasoning is required to answer natural language questions. Experimental results on these datasets demonstrate that Aqua-QA outperforms existing methods, showcasing its effectiveness in handling complex reasoning tasks in knowledge-grounded QA settings.

EvoWiki: Evaluating LLMs on Evolving Knowledge

Wei Tang, Yixin Cao, Yang Deng, Jiahao Ying, Bo Wang, Yizhe Yang, Yuyue Zhao, Qi Zhang, Xuanjing Huang, Yu-Gang Jiang, Yong Liao

Knowledge utilization is a critical aspect of LLMs, and understanding how they adapt to evolving knowledge is essential for their effective deployment. However, existing benchmarks are predominantly static, failing to capture the evolving nature of LLMs and knowledge, leading to inaccuracies and vulnerabilities such as contamination. In this paper, we introduce EvoWiki, an evolving dataset designed to reflect knowledge evolution by categorizing information into stable, evolved, and uncharted states. EvoWiki is fully auto-updatable, enabling precise evaluation of continuously changing knowledge and newly released LLMs. Through experiments with Retrieval-Augmented Generation (RAG) and Continual Learning (CL), we evaluate how effectively LLMs adapt to evolving knowledge. Our results indicate that current models often struggle with evolved knowledge, frequently providing outdated or incorrect responses. Moreover, the dataset highlights a synergistic effect between RAG and CL, demonstrating their potential to better adapt to evolving knowledge. EvoWiki provides a robust benchmark for advancing future research on the knowledge evolution capabilities of large language models.

LongDocURL: a Comprehensive Multimodal Long Document Benchmark Integrating Understanding, Reasoning, and Locating

CHAO DENG, Jiale Yuan, Pi Bu, Peijie Wang, Zhong-Zhi Li, Jian Xu, Xiao-Hui Li, Yuan Gao, Jun Song, Bo Zheng, Cheng-Lin Liu

Large vision language models (LVLMs) have improved the document understanding capabilities remarkably, enabling the handling of complex document elements, longer contexts, and a wider range of tasks. However, existing document understanding benchmarks have been limited to handling only a small number of pages and fail to provide a comprehensive analysis of layout elements locating. In this paper, we first define three primary task categories: Long Document Understanding, numerical Reasoning, and cross-element Locating, and then propose a comprehensive benchmark---LongDocURL---integrating above three primary tasks and comprising 20 sub-tasks categorized based on different primary tasks and answer evidences. Furthermore, we develop a semi-automated construction pipeline and collect 2,325 high-quality question-answering pairs, covering more than 33,000 pages of documents, significantly outperforming existing benchmarks. Subsequently, we conduct comprehensive evaluation experiments on both open-source and closed- source models across 26 different configurations, revealing critical performance gaps in this field. The code and data: https://github.com/dengc2023/LongDocURL.

TARGA: Targeted Synthetic Data Generation for Practical Reasoning over Structured Data

Xiang Huang, Jiayu Shen, Shanshan Huang, Sitao Cheng, Xiaxia Wang, Yuzhong Qu

Semantic parsing, which converts natural language queries into logic forms, plays a crucial role in reasoning within structured environments. However, existing methods encounter two significant challenges: reliance on extensive manually annotated datasets and limited generalization capability to unseen examples. To tackle these issues, we propose Targeted Synthetic Data Generation (Targa), a practical framework that dynamically generates high-relevance synthetic data without manual annotation. Starting from the pertinent entity and relation of a given question, we probe for the potential relevant queries through layer-wise expansion and cross-layer combination. Then, we generate corresponding natural language questions for these constructed queries to jointly serve as the synthetic demonstration for in-context learning. Experiments on multiple knowledge-based question answering (KBQA) datasets demonstrate that Targa, using only a 7B-parameter model, substantially outperforms existing non-fine-tuned methods that utilize close-sourced model, achieving notable improvements in F1 scores on GrailQA(+7.7) and KBQA-Agent(+12.2). Furthermore, Targa also exhibits superior sample efficiency, robustness, and generalization capabilities under non-I.I.D. settings.

Learning to Reason from Feedback at Test-Time

Yanyang Li, Michael Lyu, Liwei Wang

Solving complex tasks in a single attempt is challenging for large language models (LLMs). Iterative interaction with the environment and feedback is often required to achieve success, making effective feedback utilization a critical topic. Existing approaches either struggle with length generalization or rely on naive retries without leveraging prior information. In this paper, we introduce FTTT, a novel paradigm that formulates feedback utilization as an optimization problem at test time. Additionally, we propose a learnable test-time optimizer, OpTune, to effectively exploit feedback. Experiments on two LLMs across four reasoning datasets demonstrate that FTTT and OpTune achieve superior scalability and performance.

Pattern Recognition or Medical Knowledge? The Problem with Multiple-Choice Questions in Medicine

Maxime Griot, Jean Vanderdonckt, Demet YUKSEL, Coralie Hemptinne

Large Language Models (LLMs) such as ChatGPT demonstrate significant potential in the medical domain and are often evaluated using multiple-choice questions (MCQs) modeled on exams like the USMLE. However, such benchmarks may overestimate true clinical understanding by rewarding pattern recognition and test-taking heuristics. To investigate this, we created a fictional medical benchmark centered on an imaginary organ, the Glianorex, allowing us to separate memorized knowledge from reasoning ability. We generated textbooks and MCQs in English and French using leading LLMs, then evaluated proprietary, open-source, and domain-specific models in a zero-shot setting. Despite the fictional content, models achieved an average score of 64%, while physicians scored only 27%. Fine-tuned medical models outperformed base models in English but not in French. Ablation and interpretability analyses revealed that models frequently relied on shallow cues, test-taking strategies, and hallucinated reasoning to identify the correct choice. These results suggest that standard MCQ-based evaluations may not effectively measure clinical reasoning and highlight the need for more robust, clinically meaningful assessment methods for LLMs.

NeuSym-RAG: Hybrid Neural Symbolic Retrieval with Multiview Structuring for PDF Question Answering

Ruisheng Cao, Hanchong Zhang, Tiancheng Huang, Zhangyi Kang, Yuxin Zhang, Liangtai Sun, Hanqi Li, Yuxun Miao, Shuai Fan, Lu Chen, Kai Yu

The increasing number of academic papers poses significant challenges for researchers to efficiently acquire key details. While retrieval augmented generation (RAG) shows great promise in large language model (LLM) based automated question answering, previous works often isolate neural and symbolic retrieval despite their complementary strengths. Moreover, conventional single-view chunking neglects the rich structure and layout of PDFs, e.g., sections and tables. In this work, we propose NeuSym-RAG, a hybrid neural symbolic retrieval framework which combines both paradigms in an interactive process. By leveraging multi-view chunking and schema-based parsing, NeuSym-RAG organizes semi-structured PDF content into both the relational database and vectorstore, enabling LLM agents to iteratively gather context until sufficient to generate answers. Experiments on three full PDF-based QA datasets, including a self-annotated one AirQA-Real, show that NeuSym-RAG stably defeats both the vector-based RAG and various structured baselines, highlighting its capacity to unify both retrieval schemes and utilize multiple views.

Adaptive Retrieval Without Self-Knowledge? Bringing Uncertainty Back Home

Viktor Moskvoretskii, Maria Marina, Mikhail Salnikov, Nikolay Ivanov, Sergey Pletenev, Daria Galimzianova, Nikita Krayko, Vasily Konovalov, Irina Nikishina, Alexander Panchenko

Retrieval Augmented Generation (RAG) improves correctness of Question Answering (QA) and addresses hallucinations in Large Language Models (LLMs), yet greatly increase computational costs. Besides, RAG is not always needed as may introduce irrelevant information. Recent adaptive retrieval methods integrate LLMs' intrinsic knowledge with external information appealing to LLM self-knowledge, but they often neglect efficiency evaluations and comparisons with uncertainty estimation techniques. We bridge this gap by conducting a comprehensive analysis of 35 adaptive retrieval methods, including 8 recent approaches and 27 uncertainty estimation techniques, across 6 datasets using 10 metrics for QA performance, self-knowledge, and efficiency. Our findings show that uncertainty estimation techniques often outperform complex pipelines in terms of efficiency and self-knowledge, while maintaining comparable QA performance.

When to Speak, When to Abstain: Contrastive Decoding with Abstention

Hyuhng Joon Kim, Youna Kim, Sang-goo Lee, Taeuk Kim

Large Language Models (LLMs) demonstrate exceptional performance across diverse tasks by leveraging pre-trained (i.e., parametric) and external (i.e., contextual) knowledge. While substantial efforts have been made to enhance the utilization of both forms of knowledge, situations in which models lack relevant information remain underexplored. To investigate this challenge, we first present a controlled testbed featuring four distinct knowledge access scenarios, including the aforementioned edge case, revealing that conventional LLM usage exhibits insufficient robustness in handling all instances. Addressing this limitation, we propose Contrastive Decoding with Abstention (CDA), a novel training-free decoding method that allows LLMs to generate responses when relevant knowledge is available and to abstain otherwise. CDA estimates the relevance of both knowledge sources for a given input, adaptively deciding which type of information to prioritize and which to exclude. Through extensive experiments, we demonstrate that CDA can effectively perform accurate generation and abstention simultaneously, enhancing reliability and preserving user trust.

WebWalker: Benchmarking LLMs in Web Traversal

Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, Fei Huang

Retrieval-augmented generation (RAG) demonstrates remarkable performance across tasks in open-domain question-answering. However, traditional search engines may retrieve shallow content, limiting the ability of LLMs to handle complex, multi-layered information. To address this, we introduce WebWalkerQA, a benchmark designed to assess the ability of LLMs to perform web traversal. It evaluates the capacity of LLMs to traverse a website's subpages to extract high-quality data systematically. We propose WebWalker, which is a multi-agent framework that mimics human-like web navigation through an explore-critic paradigm. Extensive experimental results show that WebWalkerQA is challenging and demonstrates the effectiveness of RAG combined with WebWalker, through this horizontal and vertical integration in real-world scenarios.

How does Misinformation Affect Large Language Model Behaviors and Preferences?

Miao Peng, Nuo Chen, Jianheng Tang, Jia Li

Large Language Models (LLMs) have shown remarkable capabilities in knowledge-intensive tasks, while they remain vulnerable when encountering misinformation. Existing studies have explored the role of LLMs in combating misinformation, but there is still a lack of fine-grained analysis on the specific aspects and extent to which LLMs are influenced by misinformation. To bridge this gap, we present MisBench, the current largest and most comprehensive benchmark for evaluating LLMs' behavior and knowledge preference toward misinformation. MisBench consists of 10,346,712 pieces of misinformation, which uniquely considers both knowledge-based conflicts and stylistic variations in misinformation. Empirical results reveal that while LLMs demonstrate comparable abilities in discerning misinformation, they still remain susceptible to knowledge conflicts and stylistic variations. Based on these findings, we further propose a novel approach called Reconstruct to Discriminate (RtD) to strengthen LLMs' ability to detect misinformation. Our study provides valuable insights into LLMs' interactions with misinformation, and we believe MisBench can serve as an effective benchmark for evaluating LLM-based detectors and enhancing their reliability in real-world applications. Codes and data are available at: https://github.com/GKNL/MisBench.

HSCR: Hierarchical Self-Contrastive Rewarding for Aligning Medical Vision Language Models

Songtao Jiang, Yan Zhang, Yeying Jin, Zhihang Tang, Yangyang Wu, YANG FENG, Jian Wu, Zuozhu Liu

Medical Vision-Language Models (Med-VLMs) have achieved success across various tasks, yet most existing methods overlook the modality misalignment issue that can lead to untrustworthy responses in clinical settings. In this paper, we propose Hierarchical Self-Contrastive Rewarding (HSCR), a novel approach that addresses two critical challenges in Med-VLM alignment: 1) Cost-effective generation of high-quality preference data; 2) Capturing nuanced and context-aware preferences for improved alignment. HSCR first leverages the inherent capability of Med-VLMs to generate dispreferred responses with higher sampling probability. By analyzing output logit shifts after visual token dropout, we identify modality-coupled tokens that induce misalignment and derive an implicit alignment reward function. This function guides token replacement with hallucinated ones during decoding, producing high-quality dispreferred data. Furthermore, HSCR introduces a multi-level preference optimization strategy, which extends beyond traditional adjacent-level optimization by incorporating nuanced implicit preferences, leveraging relative quality in dispreferred data to capture subtle alignment cues for more precise and context-aware optimization. Extensive experiments across multiple medical tasks, including Med-VQA, medical image captioning and instruction following, demonstrate that HSCR not only enhances zero-shot performance but also significantly improves modality alignment and trustworthiness with just 2,000 training entries. Code is released on https://github.com/jiangsongtao/HSCR.

DCG-SQL: Enhancing In-Context Learning for Text-to-SQL with Deep Contextual Schema Link Graph

Jihyung Lee, Jin-Seop Lee, Jaehoon Lee, YunSeok Choi, Jee-Hyong Lee

Text-to-SQL, which translates a natural language question into an SQL query, has advanced with in-context learning of Large Language Models (LLMs). However, existing methods show little improvement in performance compared to randomly chosen demonstrations, and significant performance drops when smaller LLMs (e.g., Llama 3.1-8B) are used. This indicates that these methods heavily rely on the intrinsic capabilities of hyper-scaled LLMs, rather than effectively retrieving useful demonstrations. In this paper, we propose a novel approach for effectively retrieving demonstrations and generating SQL queries. We construct a Deep Contextual Schema Link Graph, which contains key information and semantic relationship between a question and its database schema items. This graph-based structure enables effective representation of Text-to-SQL samples and retrieval of useful demonstrations for in-context learning. Experimental results on the Spider benchmark demonstrate the effectiveness of our approach, showing consistent improvements in SQL generation performance and efficiency across both hyper-scaled LLMs and small LLMs. The code is available at https://github.com/jjklle/DCG-SQL.

Improving Contextual Faithfulness of Large Language Models via Retrieval Heads-Induced Optimization

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuchun Fan, Xiachong Feng, Yangfan Ye, Weihong Zhong, Yuxuan Gu, Baoxin Wang, Dayong Wu, Guoping Hu, Bing Qin

Ensuring contextual faithfulness in retrieval-augmented large language models (LLMs) is crucial for building trustworthy information-seeking systems, particularly in long-form question-answering (LFQA) scenarios. In this work, we identify a salient correlation between LFQA faithfulness and retrieval heads, a set of attention heads responsible for retrieving contextual information. Leveraging this insight, we propose RHIO, a framework designed to teach LLMs to explicitly discriminate between faithful and unfaithful generations. RHIO first augments unfaithful samples that simulate realistic model-intrinsic errors by selectively masking retrieval heads. Then, these samples are incorporated into joint training, enabling the model to distinguish unfaithful outputs from faithful ones conditioned on control tokens. Furthermore, these control tokens are leveraged to self-induce contrastive outputs, amplifying their difference through contrastive decoding. Additionally, to facilitate the evaluation of contextual faithfulness, we also introduce GroundBench, a comprehensive benchmark compiled from five existing LFQA datasets. Extensive experimental results on GroundBench demonstrate that RHIO significantly improves faithfulness, even outperforming GPT-4o.

Can LLMs Evaluate Complex Attribution in QA? Automatic Benchmarking using Knowledge Graphs

Nan Hu, Jiaoyan Chen, Yike Wu, Guilin Qi, Hongru WANG, Sheng Bi, Yongrui Chen, Tongtong Wu, Jeff Z. Pan

Attributed Question Answering (AQA) has attracted wide attention, but there are still several limitations in evaluating the attributions, including lacking fine-grained attribution categories, relying on manual annotations, and failing to compare attributions with only subtle differences. To bridge these gaps, we introduce Complex Attributed Question Answering (CAQA), a large-scale benchmark containing comprehensive attribution categories, automatically generated using Knowledge Graphs (KGs), and complex attribution scenarios. We have conducted extensive experiments to verify the effectiveness of CAQA, including the benchmarking of 25 automatic evaluators, their comparison with human evaluators, the testing of LLM evaluators fine-tuned by CAQA and so on. These experiments also lead to a series of important findings that can benefit the future research of AQA.

RUBY: An Effective Framework for Multi-Constraint Multi-Hop Question Generation

Wenzhuo Zhao, Shuangyin Li

Inspired by theories in language psychology, it is natural to consider more constraints, such as intentions, logic, knowledge, etc., when a complex or multi-hop question is generated. As the subtask of Multi-Hop Question Generation (MHQG), the task of Multi-Constraint Multi-Hop Question Generation (MCHQG) is more aligned with human question theories. However, it is hard to determine how to bring various high-dimensional semantic constraints, and how to integrate each constraint across all hops when a multi-hop question is being generating. To address these challenges, we introduce an effective framework which includes constraint dimensionality reduction and divide-and-conquer-based dynamic projection; we call it RUBY. The proposed RUBY contains a module of high-dimensional semantic constraint dimension reduction and a module of sub-question answer pairs-based multi-hop question generation. Meanwhile, a Reasoning Dynamic Projection strategy is tailored to effectively incorporate the constraints into every hop of the multi-hop question. The experimental results demonstrate that RUBY consistently outperforms baseline models, which suggest that RUBY is able to effectively capture and integrate semantic constraints, leading to more accurate and human-like multi-hop question generation. Our code and data are available.

RARE: Retrieval-Augmented Reasoning Enhancement for Large Language Models

Hieu Tran, Zonghai Yao, Zhichao Yang, Junda Wang, Yifan Zhang, Feiyun Ouyang, Shuo Han, hong yu

This work introduces RARE (Retrieval-Augmented Reasoning Enhancement), a versatile extension to the mutual reasoning framework (rStar), aimed at enhancing reasoning accuracy and factual integrity across large language models (LLMs) for complex, knowledge-intensive tasks such as medical and commonsense reasoning. RARE incorporates two innovative actions within the Monte Carlo Tree Search (MCTS) framework: (A6), which generates search queries based on the initial problem statement, performs information retrieval using those queries, and augments reasoning with the retrieved data to formulate the final answer; and (A7), which leverages information retrieval specifically for generated sub-questions and re-answers these sub-questions with the relevant contextual information. Additionally, a Retrieval-Augmented Factuality Scorer is proposed to replace the original discriminator, prioritizing reasoning paths that meet high standards of factuality. Experimental results with LLaMA 3.1 show that RARE enables open-source LLMs to achieve competitive performance with top closed-source models like GPT-4 and GPT-4o. This research establishes RARE as a scalable solution for improving LLMs in domains where logical coherence and factual integrity are critical.

Micro-Act: Mitigate Knowledge Conflict in Question Answering via Actionable Self-Reasoning

Nan Huo, Jinyang Li, Bowen Qin, Ge Qu, Xiaolong Li, Xiaodong Li, Chenhao Ma, Reynold Cheng

Retrieval-Augmented Generation (RAG) systems commonly suffer from **Knowledge Conflicts**, where retrieved external knowledge contradicts the inherent, parametric knowledge of large language models (LLMs). It adversely affects performance on downstream tasks such as question answering (QA). Existing approaches often attempt to mitigate conflicts by directly comparing two knowledge sources in a side-by-side manner, but this can overwhelm LLMs with extraneous or lengthy contexts, ultimately hindering their ability to identify and mitigate inconsistencies. To address this issue, we propose **Micro-Act** a framework with a hierarchical action space that automatically perceives context complexity and adaptively decomposes each knowledge source into a sequence of fine-grained comparisons. These comparisons are represented as actionable steps, enabling reasoning beyond the superficial context. Through extensive experiments on five benchmark datasets, Micro-Act consistently achieves significant increase in QA accuracy over state-of-the-art baselines across all 5 datasets and 3 conflict types, especially in temporal and semantic types where all baselines fail significantly. More importantly, Micro-Act exhibits robust performance on non-conflict questions simultaneously, highlighting its practical value in real-world RAG applications.

Is That Your Final Answer? Test-Time Scaling Improves Selective Question Answering

William Jurayj, Jeffrey Cheng, Benjamin Van Durme

Scaling the test-time compute of large language models has demonstrated impressive performance on reasoning benchmarks. However, existing evaluations of test-time scaling make the strong assumption that a reasoning system should always give an answer to any question provided. This overlooks concerns about whether a model is confident in its answer, and whether it is appropriate to always provide a response. To address these concerns, we extract confidence scores during reasoning for thresholding model responses. We find that increasing compute budget at inference time not only helps models answer more questions correctly, but also increases confidence in correct responses. We then extend the current paradigm of zero-risk responses during evaluation by considering settings with non-zero levels of response risk, and suggest a recipe for reporting evaluations under these settings.

Beyond Prompting: An Efficient Embedding Framework for Open-Domain Question Answering

Zhanghao Hu, Hanqi Yan, Qinglin Zhu, Zhenyi Shen, Yulan He, Lin Gui

Large language models (LLMs) have recently pushed open-domain question answering (ODQA) to new frontiers. However, prevailing retriever–reader pipelines often depend on multiple rounds of prompt-level instructions, leading to high computational overhead, instability, and suboptimal retrieval coverage. In this paper, we propose EmbQA, an embedding-level framework that alleviates these shortcomings by enhancing both the retriever and the reader. Specifically, we refine query representations via lightweight linear layers under an unsupervised contrastive learning objective, thereby reordering retrieved passages to highlight those most likely to contain correct answers. Additionally, we introduce an exploratory embedding that broadens the model's latent semantic space to diversify candidate generation and employs an entropy-based selection mechanism to choose the most confident answer automatically. Extensive experiments across three open-source LLMs, three retrieval methods, and four ODQA benchmarks demonstrate that EmbQA substantially outperforms recent baselines in both accuracy and efficiency.

Knowledge Image Matters: Improving Knowledge-Based Visual Reasoning with Multi-Image Large Language Models

Guanghui Ye, Huan Zhao, Zhixue Zhao, Xupeng Zha, Yang Liu, Zhihua Jiang

We revisit knowledge-based visual reasoning (KB-VR) in light of modern advances in multimodal large language models (MLLMs), and make the following contributions: (i) We propose Visual Knowledge Card (VKC) – a novel image that incorporates not only internal visual knowledge (e.g., scene-aware information) detected from the raw image, but also external world knowledge (e.g., attribute or object knowledge) produced by a knowledge generator; (ii) We present VKC-based Multi-Image Reasoning (VKC-MIR) – a four-stage pipeline which harnesses a state-of-the-art scene perception engine to construct an initial VKC (Stage-1), a powerful LLM to generate relevant domain knowledge (Stage-2), an excellent image editing toolkit to introduce generated knowledge into the updated VKC (Stage-3), and finally, an emerging multi-image MLLM to solve the VKC-enhanced task (Stage-4). By performing experiments on three popular KB-VR benchmarks, our approach achieves new state-of-the-art results compared to previous top-performing models.

Not All Terms Matter: Recall-Oriented Adaptive Learning for PLM-aided Query Expansion in Open-Domain Question Answering

Xinran Chen, Ben He, Xuanang Chen, Le Sun

The effectiveness of open-domain question answering (ODQA), particularly those employing a retriever-reader architecture, depends on the ability to recall relevant documents - a critical step that enables the reader to accurately extract answers. To enhance this retrieval phase, current query expansion (QE) techniques leverage pre-trained language models (PLM) to mitigate word mismatches and improve the recall of relevant documents. Despite their advancements, these techniques often treat all expanded terms uniformly, which can lead to less-than-optimal retrieval outcomes. In response, we propose a novel Recall-oriented Adaptive Learning (ReAL) method, which iteratively adjusts the importance weights of QE terms based on their relevance, thereby refining term distinction and enhancing the separation of relevant terms. Specifically, ReAL employs a similarity-based model to classify documents into pseudo-relevant and pseudo-irrelevant sets, and then optimizes term weights via two tailored loss functions to maximize the scoring gap between them. Experiments on four ODQA datasets and five QE methods show that ReAL consistently enhances retrieval accuracy and overall end-to-end QA performance, providing a robust and efficient solution for improving QE strategies in ODQA scenarios.

Decoding on Graphs: Faithful and Sound Reasoning on Knowledge Graphs through Generation of Well-Formed Chains

Kun LI, Tianhua Zhang, Xixin Wu, Hongyin Luo, James R. Glass, Helen M. Meng

Knowledge Graphs (KGs) can serve as reliable knowledge sources for question answering (QA) due to their structured representation of knowledge. Existing research on the utilization of KG for large language models (LLMs) prevalently relies on subgraph retriever or iterative prompting, overlooking the potential synergy of LLMs' step-wise reasoning capabilities and KGs' structural nature. In this paper, we present DoG (Decoding on Graph), a novel framework that facilitates a deep synergy between LLMs and KGs. We first define a concept, well-formed chain, which consists of a sequence of interrelated fact triplets on the KGs, starting from question entities and leading to answers. We argue that this concept can serve as a principle for making faithful and sound reasoning for KGQA. To enable LLMs to generate well-formed chains, we propose graph-aware constrained decoding, in which a constraint derived from the topology of the KG regulates the decoding process of the LLMs. This constrained decoding method ensures the generation of well-formed chains while making full use of the step-wise reasoning capabilities of LLMs. Based on the above, DoG, a training-free approach, is able to provide faithful and sound reasoning trajectories grounded on the KGs. Experiments across various KGQA tasks with different background KGs demonstrate that DoG achieves superior and robust performance. DoG also shows general applicability with various open-source LLMs.

Deliberate Reasoning for Language Models as Structure-aware Planning with Accurate World Model

Siheng Xiong, Ali Payani, Yuan Yang, Faramarz Fekri

Enhancing the reasoning capabilities of language models (LMs) remains a key challenge, especially for tasks that require complex, multi-step decision-making where existing Chain-of-Thought (CoT) approaches struggle with consistency and verification. In this paper, we propose a novel reasoning framework, referred to as Structure-aware Planning with an Accurate World Model (SWAP), that integrates structured knowledge representation with learned planning. Unlike prior methods that rely purely on natural language reasoning, SWAP leverages entailment graphs to encode structured dependencies and enable symbolic verification of intermediate steps. To systematically construct and update the graph, SWAP employs a policy model to propose candidate expansions and a world model to predict structural updates. To improve accuracy, the world model generates multiple alternative updates, and a discriminator re-ranks them based on plausibility. To encourage diverse exploration, we introduce Diversity-based Modelling (DM), which samples candidates from the remaining probability mass after removing previously sampled candidates from the original policy distribution. Additionally, SWAP improves the discrimination accuracy through Contrastive Ranking (CR), which directly compares candidates within prompts and incorporates meta-knowledge to improve ranking quality. We evaluate SWAP across diverse reasoning-intensive benchmarks including math reasoning, logical reasoning, and coding tasks. Extensive experiments demonstrate that SWAP significantly improves upon the base models and consistently outperforms existing reasoning methods.

Logic-Regularized Verifier Elicits Reasoning from LLMs

Xinyu Wang, Changzhi Sun, Lian Cheng, Yuanbin Wu, Dell Zhang, Xuelong Li, Xiaoling Wang

Verifiers are crucial components for enhancing modern LLMs' reasoning capability. Typical verifiers require resource-intensive supervised dataset construction, which is costly and faces limitations in data diversity. In this paper, we propose LOVER, an unsupervised verifier regularized by logical rules. LOVER treats the verifier as a binary latent variable, utilizing internal activations and enforcing three logical constraints on multiple reasoning paths: negation consistency, intra-group consistency, and inter-group consistency (grouped by the final answer). By incorporating logical rules as priors, LOVER can leverage unlabeled examples and is directly compatible with any off-the-shelf LLMs. Experiments on 10 datasets demonstrate that LOVER significantly outperforms unsupervised baselines, achieving performance comparable to the supervised verifier (reaching its 95% level on average).

CiteEval: Principle-Driven Citation Evaluation for Source Attribution

Yumo Xu, Peng Qi, Jifan Chen, Kunlun Liu, Rujun Han, Lan Liu, Bonan Min, Vittorio Castelli, Arshit Gupta, Zhiguo Wang

Citation quality is crucial in information-seeking systems, directly influencing trust and the effectiveness of information access. Current evaluation frameworks, both human and automatic, mainly rely on Natural Language Inference (NLI) to assess binary or ternary supportiveness from cited sources, which we argue is a suboptimal proxy for citation evaluation. In this work we introduce CiteEval, a citation evaluation framework driven by principles focusing on fine-grained citation assessment within a broad context, encompassing not only the cited sources but the full retrieval context, user query, and generated text. Guided by the proposed framework, we construct CiteBench, a multi-domain benchmark with high-quality human annotations on citation quality. To enable efficient evaluation, we further develop CiteEval-Auto, a suite of model-based metrics that exhibit strong correlation with human judgments. Experiments across diverse systems demonstrate CiteEval-Auto's superior ability to capture the multifaceted nature of citations compared to existing metrics, offering a principled and scalable approach to evaluate and improve model-generated citations.

Session 7: Resources and Evaluation (Tuesday 29 July, 10:30-12:00, Hall 4/5)

MinosEval: Distinguishing Factoid and Non-Factoid for Tailored Open-Ended QA Evaluation with LLMs

Yongqi Fan, Yating Wang, Guandong Wang, Zhai Jie, Jingping Liu, Qi Ye, Tong Ruan

Open-ended question answering (QA) is a key task for evaluating the capabilities of large language models (LLMs). Compared to closed-ended QA, it demands longer answer statements, more nuanced reasoning processes, and diverse expressions, making refined and interpretable automatic evaluation both crucial and challenging. Traditional metrics like ROUGE and BERTScore struggle to capture semantic similarities due to different patterns between model responses and reference answers. Current LLM-based evaluation approaches, such as pairwise or listwise comparisons of candidate answers, lack intuitive interpretability. While pointwise scoring of each response provides some descriptions, it fails to adapt across different question contents. Most notably, existing methods overlook the distinction between factoid and non-factoid questions. To address these challenges, we propose \textbf{MinosEval}, a novel evaluation method that first distinguishes open-ended questions and then ranks candidate answers using different evaluation strategies. For factoid questions, it applies an adaptive key-point scoring strategy, while for non-factoid questions, it uses an instance-aware listwise ranking strategy. Experiments on multiple open-ended QA datasets, including self-built ones with more candidate responses to complement community resources, show that MinosEval better aligns with human annotations and offers more interpretable results.

CoDet-M4: Detecting Machine-Generated Code in Multi-Lingual, Multi-Generator and Multi-Domain Settings

Daniil Orel, Dilshod Azizov, Preslav Nakov

Large language models (LLMs) have revolutionized code generation, automating programming with remarkable efficiency. However, these advancements challenge programming skills, ethics, and assessment integrity, making the detection of LLM-generated code essential for maintaining accountability and standards. While, there has been some research on this problem, it generally lacks domain coverage and robustness, and only covers a small number of programming languages. To this end, we propose a framework capable of distinguishing between human- and LLM-written code across multiple programming languages, code generators, and domains. We use a large-scale dataset from renowned platforms and LLM-based code generators, alongside applying rigorous data quality checks, feature engineering, and comparative analysis using evaluation of traditional machine learning models, pre-trained language models (PLMs), and LLMs for code detection. We perform an evaluation on out-of-domain scenarios, such as detecting the authorship and hybrid authorship of generated code and generalizing to unseen models, domains, and programming languages. Moreover, our extensive experiments show that our framework effectively distinguishes human- from LLM-written code and sets a new benchmark for this task.

taz2024full: Analysing German Newspapers for Gender Bias and Discrimination across Decades

Stefanie Urchs, Veronika Thurner, Matthias Aßenmacher, Christian Heumann, Stephanie Thiemichen

Open-access corpora are essential for advancing natural language processing (NLP) and computational social science (CSS). However, large-scale resources for German remain limited, restricting research on linguistic trends and societal issues such as gender bias. We present taz2024full, the largest publicly available corpus of German newspaper articles to date, comprising over 1.8 million texts from taz, spanning 1980 to 2024. As a demonstration of the corpus's utility for bias and discrimination research, we analyse gender representation across four decades of reporting. We find a consistent overrepresentation of men, but also a gradual shift toward more balanced coverage in recent years. Using a scalable, structured analysis pipeline, we provide a foundation for studying actor mentions, sentiment, and linguistic framing in German journalistic texts. The corpus supports a wide range of applications, from diachronic language analysis to critical media studies, and is freely available to foster inclusive and reproducible research in German-language NLP.

LCFO: Long Context and Long Form Output Dataset and Benchmarking

Marta R. Costa-jussà, Pierre Andrews, Mariano Coria Meglioli, Joy Chen, Joe Chuang, David Dale, Christophe Ropers, Alexandre Mourachko, Eduardo Sánchez, Holger Schwenk, Tuan A. Tran, Arina Turkatenko, Carleigh Wood

This paper presents the Long Context and Form Output (LCFO) benchmark, a novel evaluation framework for assessing gradual summarization and summary expansion capabilities across diverse domains. LCFO consists of long input documents (5k words average length), each of which comes with three summaries of different lengths (20%, 10%, and 5% of the input text), as well as approximately 15 questions and answers (QA) related to the input content. Notably, LCFO also provides alignments between specific QA pairs and corresponding summaries in 7 domains. The primary motivation behind providing summaries of different lengths is to establish a controllable framework for generating long texts from shorter inputs, i.e. summary expansion. %, enabling a more nuanced evaluation of summary expansion capabilities. To establish an evaluation metric framework for summarization and summary expansion, we provide human evaluation scores for human-generated outputs, as well as results from various state-of-the-art large language models (LLMs). GPT-4o-mini achieves best human scores among automatic systems in both summarization and summary expansion tasks (+10% and +20%, respectively). It even surpasses human output quality in the case of short summaries (+7%). Overall automatic metrics achieve low correlations with human evaluation scores (0.4) but moderate correlation on specific evaluation aspects such as fluency and attribution (0.6).

CoMuMDR: Code-mixed Multi-modal Multi-domain corpus for Discourse paRsing in conversations

Divyaksh Shukla, Ritesh Baviskar, Dwijesh Gohil, Aniket Tiwari, Atul Shree, Ashutosh Modi

Discourse parsing is an important task useful for NLU applications such as summarization, machine comprehension, and emotion recognition. The current discourse parsing datasets based on conversations consists of written English dialogues restricted to a single domain. In this resource paper, we introduce CoMuMDR: Code-mixed Multi-modal Multi-domain corpus for Discourse paRsing in conversations. The corpus (code-mixed in Hindi and English) has both audio and transcribed text and is annotated with nine discourse relations. We experiment with various SoTA baseline models; the poor performance of SoTA models highlights the challenges of multi-domain code-mixed data, pointing towards the need for developing better models for such realistic settings.

2M-BELEBELE: Highly Multilingual Speech and American Sign Language Comprehension Dataset Download PDF

Marta R. Costa-jussà, Bokai YU, Pierre Andrews, Belen Alastruey, Necati Cihan Camgoz, Joe Chuang, Jean Maillard, Christophe Ropers, Arina Turkatenko, Carleigh Wood

We introduce the first highly multilingual speech and American Sign Language (ASL) comprehension dataset by extending BELEBELE. Our dataset covers 74 spoken languages at the intersection of BELEBELE and FLEURS, and one sign language (ASL). We evaluate 2M-BELEBELE dataset for both 5-shot and zero-shot settings and across languages, the speech comprehension accuracy is 10% average lower compared to reading comprehension.

WirelessMathBench: A Mathematical Modeling Benchmark for LLMs in Wireless Communications

Xin Li, Mengbing Liu, LI WEI, Jiancheng An, Merouane Abdelkader DEBBAH, Chau Yuen
Large Language Models (LLMs) have achieved impressive results across a broad array of tasks, yet their capacity for complex, domain-specific mathematical reasoning—particularly in wireless communications—remains underexplored. In this work, we introduce WirelessMathBench, a novel benchmark specifically designed to evaluate LLMs on mathematical modeling challenges to wireless communications engineering. Our benchmark consists of 587 meticulously curated questions sourced from 40 state-of-the-art research papers, encompassing a diverse spectrum of tasks ranging from basic multiple-choice questions to complex equation completion tasks, including both partial and full completions, all of which rigorously adhere to physical and dimensional constraints. Through extensive experimentation with leading LLMs, we observe that while many models excel in basic recall tasks, their performance degrades significantly when reconstructing partially or fully obscured equations, exposing fundamental limitations in current LLMs. Even DeepSeek-R1, the best performer on our benchmark, achieves an average accuracy of only 38.05%, with a mere 7.83% success rate in full equation completion. By publicly releasing WirelessMathBench along with the evaluation toolkit, we aim to advance the development of more robust, domain-aware LLMs for wireless system analysis and broader engineering applications.

DEMO: Reframing Dialogue Interaction with Fine-grained Element Modeling

Minzheng Wang, Xinghua Zhang, Kun Chen, Nan Xu, Haiyang Yu, Fei Huang, Wenji Mao, Yongbin Li
Large language models (LLMs) enabled dialogue systems have become one of the central modes in human-machine interaction, which bring about vast amounts of conversation logs and increasing demand for dialogue generation. The dialogue's life-cycle spans from Prelude through Interlocution to Epilogue, encompassing rich dialogue elements. Despite large volumes of dialogue-related studies, there is a lack of systematic investigation into the dialogue stages to frame benchmark construction that covers comprehensive dialogue elements. This hinders the precise modeling, generation and assessment of LLMs-based dialogue systems. To bridge this gap, in this paper, we introduce a new research task—Dialogue Element MOdeling, including Element Awareness and Dialogue Agent Interaction and propose a novel benchmark, DEMO, designed for a comprehensive dialogue modeling and assessment. On this basis, we further build the DEMO agent with the adept ability to model dialogue elements via imitation learning. Extensive experiments on DEMO indicate that current representative LLMs still have considerable potential for enhancement, and our DEMO agent performs well in both dialogue element modeling and out-of-domain tasks.

Physics: Benchmarking Foundation Models for PhD-Qualifying Exam Physics Problem Solving

Kaiyue Feng, Yilun Zhao, Yixin Liu, Tianyu Yang, Chen Zhao, John Sous, Arman Cohan
We introduce Physics, a comprehensive benchmark for PhD-qualifying exam physics problem solving. It contains 1297 expert-annotated problems covering six core areas: classical mechanics, quantum mechanics, thermodynamics and statistical mechanics, electromagnetism, atomic physics, and optics. Each problem requires advanced physics knowledge and mathematical reasoning. We develop a robust automated evaluation system for precise and reliable validation. Our evaluation of leading foundation models reveals substantial limitations. Even the most advanced model, o3-mini, achieves only 59.9% accuracy, highlighting significant challenges in solving high-level scientific problems. Through comprehensive error analysis, exploration of diverse prompting strategies, and RAG-based knowledge augmentation, we identify key areas for improvement, laying the foundation for future advancements.

NeoQA: Evidence-based Question Answering with Generated News Events

Max Glockner, Xiang Jiang, Leonardo F. R. Ribeiro, Iryna Gurevych, Markus Dreyer

Evaluating Retrieval-Augmented Generation (RAG) in large language models (LLMs) is challenging because benchmarks can quickly become stale. Questions initially requiring retrieval may become answerable from pretraining knowledge as newer models incorporate more recent information during pretraining, making it difficult to distinguish evidence-based reasoning from recall. We introduce NeoQA (News Events for Out-of-training Question Answering), a benchmark designed to address this issue. To construct NeoQA, we generated timelines and knowledge bases of fictional news events and entities along with news articles and Q\&A pairs to prevent LLMs from leveraging pretraining knowledge, ensuring that no prior evidence exists in their training data. We propose our dataset as a new platform for evaluating evidence-based question answering, as it requires LLMs to generate responses exclusively from retrieved evidence and only when sufficient evidence is available. NeoQA enables controlled evaluation across various evidence scenarios, including cases with missing or misleading details. Our findings indicate that LLMs struggle to distinguish subtle mismatches between questions and evidence, and suffer from short-cut reasoning when key information required to answer a question is missing from the evidence, underscoring key limitations in evidence-based reasoning.

WebNLG-IT: Construction of an aligned RDF-Italian corpus through Machine Translation techniques

Michael Oliverio, Pier Felice Balestrucci, Alessandro Mazzei, Valerio Basile

The main goal of this work is the creation of the Italian version of the WebNLG corpus through the application of Neural Machine Translation (NMT) and post-editing with hand-written rules. To achieve this goal, in a first step, several existing NMT models were analysed and compared in order to identify the system with the highest performance on the original corpus. In a second step, after using the best NMT system, we semi-automatically designed and applied a number of rules to refine and improve the quality of the produced resource, creating a new corpus named WebNLG-IT. We used this resource for fine-tuning several LLMs for RDF-to-text tasks. In this way, comparing the performance of LLM-based generators on both Italian and English, we have (1) evaluated the quality of WebNLG-IT with respect to the original English version, (2) released the first fine-tuned LLM-based system for generating Italian from semantic web triples and (3) introduced an Italian version of a modular generation pipeline for RDF-to-text.

SEA-HELM: Southeast Asian Holistic Evaluation of Language Models

Yosephine Susanto, Adithya Venkatadri Hulagadri, Jann Railey Montalan, Jian Gang Ngui, Xianbin Yong, Wei Qi Leong, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Yifan Mai, William Chandra Tjhi

With the rapid emergence of novel capabilities in Large Language Models (LLMs), the need for rigorous multilingual and multicultural benchmarks that are integrated has become more pronounced. Though existing LLM benchmarks are capable of evaluating specific capabilities of LLMs in English as well as in various mid- to low-resource languages, including those in the Southeast Asian (SEA) region, a comprehensive and authentic evaluation suite for the SEA languages has not been developed thus far. Here, we present SEA-HELM, a holistic linguistic and cultural LLM evaluation suite that emphasizes SEA languages, comprising five core pillars: (1) NLP CLASSICS, (2) LLM-SPECIFICS, (3) SEA LINGUISTICS, (4) SEA CULTURE, (5) SAFETY. SEA-HELM currently supports Filipino, Indonesian, Tamil, Thai, and Vietnamese. We also introduce the SEA-HELM leaderboard, which allows users to understand models' multilingual and multicultural performance in a systematic and user-friendly manner.

MDIT-Bench: Evaluating the Dual-Implicit Toxicity in Large Multimodal Models

Bohan Jin, Shuhan Qi, Kehai Chen, Xinyi Guo, Xuan Wang

The widespread use of Large Multimodal Models (LMMs) has raised concerns about model toxicity. However, current research mainly focuses on explicit toxicity, with less attention to some more implicit toxicity regarding prejudice and discrimination. To address this limitation, we introduce a subtler type of toxicity named dual-implicit toxicity and a novel toxicity benchmark termed MDIT-Bench: Multimodal Dual-Implicit Toxicity Benchmark. Specifically, we first create the MDIT-Dataset with dual-implicit toxicity using the proposed Multi-stage Human-in-loop In-context Generation method. Based on this dataset, we construct the MDIT-Bench, a benchmark for evaluating the sensitivity of models to dual-implicit toxicity, with 317,638 questions covering 12 categories, 23 subcategories, and 780 topics. MDIT-Bench includes three difficulty levels, and we propose a metric to measure the toxicity gap exhibited by the model across them. In the experiment, we conducted MDIT-Bench on 13 prominent LMMs, and the results show that these LMMs cannot handle dual-implicit toxicity effectively. The model's performance drops significantly in hard level, revealing that these LMMs still contain a significant amount of hidden but activatable toxicity. The data will be released upon the paper's acceptance.

Disentangling Text and Math in Word Problems: Evidence for the Bidimensional Structure of Large Language Models' Reasoning

Pedro Calais, Gabriel Franco, Zilu Tang, Themistoklis Nikas, Wagner Meira Jr., Evimaria Terzi, Mark Crovella

Do LLMs process text and mathematics as a unified skill, or do these components rely on distinct underlying mechanisms? We investigate this question by disentangling the textual interpretation and mathematical solving steps in word problems drawn from Brazil’s largest college entrance exam (ENEM) and GSM8K, a popular grade school-level benchmark. Using the symbolic solver SymPy, we transform word problems into equivalent purely mathematical representations, isolating equation formulation from textual comprehension. Our extended benchmarks enable a structured analysis of LLM performance across these two dimensions. Through empirical evaluations, we find that small-scale LLMs struggle significantly more with text interpretation than with equation solving, with accuracy dropping by a factor of 2 to 7 when solving full word problems compared to their math-only counterparts. Exploratory factor analysis confirms a bidimensional structure in LLM reasoning, where models exhibit distinct proficiencies in textual and mathematical components, underscoring the need for targeted improvements in language comprehension. By analyzing the latent factors associated with each model, our findings provide a framework for researchers and practitioners to make informed choices when selecting models based on computational costs and the nature of their tasks.

Express What You See: Can Multimodal LLMs Decode Visual Ciphers with Intuitive Semiosis Comprehension?

Jiayi Kuang, Yinghui Li, Chen Wang, Haohao Luo, Ying Shen, Wenhao Jiang

Bridging the gap between visual and language remains a pivotal challenge for multimodal community. Traditional VQA benchmarks encounter modality gap and over-reliance on language priors, whereas human cognition excels at intuitive semiosis, associating abstract visual symbols to linguistic semantics. Inspired by this neurocognitive mechanism, we focus on emojis, the visual cipher conveying abstract textual semantics. Specifically, we propose a novel task of generating abstract linguistics from emoji sequence images, where such reasoning underpins critical application in cryptography, thus challenging MLLMs' reasoning of decoding complex semantics of visual ciphers. We introduce eWe-bench (Express What you SeE), assessing MLLMs' capability of intuitive semiosis like human. Our data construction framework ensures high visual sensitivity and data quality, which can be extended to future data enhancement. Evaluation results on advanced MLLMs highlight critical deficiencies in visual intuitive symbolic reasoning. We believe our interesting insights for advancing visual semiosis in MLLMs will pave the way for cryptographic analysis and high-level intuitive cognition intelligence of MLLMs.

BioHopR: A Benchmark for Multi-Hop, Multi-Answer Reasoning in Biomedical Domain

Yunsoo Kim, Yusuf Abdulle, Honghan Wu

Biomedical reasoning often requires traversing interconnected relationships across entities such as drugs, diseases, and proteins. Despite the increasing prominence of large language models (LLMs), existing benchmarks lack the ability to evaluate multi-hop reasoning in the biomedical domain, particularly for queries involving one-to-many and many-to-many relationships. This gap leaves the critical challenges of biomedical multi-hop reasoning underexplored. To address this, we introduce BioHopR, a novel benchmark designed to evaluate multi-hop, multi-answer reasoning in structured biomedical knowledge graphs. Built from the comprehensive PrimeKG, BioHopR includes 1-hop and 2-hop reasoning tasks that reflect real-world biomedical complexities. Evaluations of state-of-the-art models reveal that O3-mini, a proprietary reasoning-focused model, achieves 37.93% precision on 1-hop tasks and 14.57% on 2-hop tasks, outperforming proprietary models such as GPT4O and open-source biomedical models including HuatuoGPT-o1-70B and Llama-3.3-70B. However, all models exhibit significant declines in multi-hop performance, underscoring the challenges of resolving implicit reasoning steps in the biomedical domain. By addressing the lack of benchmarks for multi-hop reasoning in biomedical domain, BioHopR sets a new standard for evaluating reasoning capabilities and highlights critical gaps between proprietary and open-source models while paving the way for future advancements in biomedical LLMs. BioHopR is available at https://huggingface.co/datasets/knowlab-research/BioHopR.

BriefMe: A Legal NLP Benchmark for Assisting with Legal Briefs

Jesse Woo, Fateme Hashemi Chaleshtori, Ana Marasovic, Kenneth Marino

A core part of legal work which has been under-explored in Legal NLP is the writing and editing of legal briefs. Writing and editing briefs not only requires a thorough understanding of the law of a jurisdiction from judgments to statutes, but also the ability to make new arguments to try to expand the law in a new direction and make novel and creative arguments that are persuasive to judges. To capture and evaluate these legal skills in language models, we introduce BriefMe, a new dataset focused on legal briefs. It contains three tasks for language models to assist legal professionals in writing briefs: argument summarization, argument completion, and case retrieval. In this work we describe the creation of these tasks, analyze these tasks, and show how current models perform. We see that current-generation large language models (LLMs) are already quite good at the summarization and completion tasks, even beating human-generated headings. We hope this dataset encourages more development in Legal NLP in ways that will specifically aid people in performing legal work.

SafeLawBench: Towards Safe Alignment of Large Language Models

Chuxue Cao, Han Zhu, Jiaming Ji, Qichao Sun, Zhenghao Zhu, WU YINYU, Josef Dai, Yaodong Yang, Sirui Han, Yike Guo

With the growing prevalence of large language models (LLMs), the safety of LLMs has raised significant concerns. However, evaluating their safety is still lack of definitive standards due to the subjective nature of current safety benchmarks. To address this gap, we conducted the first exploration of LLMs safety evaluation from a legal perspective by proposing the SafeLawBench benchmark. The SafeLawBench categorizes safety risks into three levels based on legal standards, providing a systematic and comprehensive framework for evaluation. It comprises 24,860 multi-choice questions and 1,107 open-domain question-answering (QA) tasks. Our evaluation encompassed 17 LLMs using both zero-shot and few-shot prompting, highlighting the advantages of closed-source models over open-source ones. We also evaluated the safety-related reasoning stability of the LLMs and their refusal behavior. Additionally, we found that a majority voting mechanism can enhance model performance. Notably, even leading SOTA models like GPT-4o and Claude-3.5-Sonnet have not exceeded 80.3% performance, while the average performance of 17 LLMs remains at 67% on SafeLawBench. We urge the community to prioritize research on the safety of LLMs.

NativQA: Multilingual Culturally-Aligned Natural Query for LLMs

Md. Arid Hasan, Maram Hasanain, Fatema Ahmad, Sahinur Rahman Laskar, Sunaya Upadhyay, Vrunda N Sukhadia, Mucahid Kutlu, Shammur Absar Chowdhury, Firoj Alam

Natural Question Answering (QA) datasets play a crucial role in evaluating the capabilities of large language models (LLMs), ensuring their effectiveness in real-world applications. Despite the numerous QA datasets that have been developed and some work has been done in parallel, there is a notable lack of a framework and large scale region-specific datasets queried by native users in their own languages. This gap hinders the effective benchmarking and the development of fine-tuned models for regional and cultural specificities. In this study, we propose a scalable, language-independent framework, NativQA, to seamlessly construct culturally and regionally aligned QA datasets in native languages, for LLM evaluation and tuning. We demonstrate the efficacy of the proposed framework by designing a multilingual natural QA dataset, MultiNativQA, consisting of 64k manually annotated QA pairs in seven languages, ranging from high to extremely low resource, based on queries from native speakers from 9 regions covering 18 topics. We benchmark open- and closed-source LLMs with the MultiNativQA dataset. We made the framework NativQA, MultiNativQA dataset, and other experimental scripts publicly available for the community (https://anonymous.com/).

AdParaphrase v2.0: Generating Attractive Ad Texts Using a Preference-Annotated Paraphrase Dataset

Soichiro Murakami, Peinan Zhang, Hidetaka Kamigaito, Hiroya Takamura, Manabu Okumura

Identifying factors that make ad text attractive is essential for advertising success. This study proposes AdParaphrase v2.0, a dataset for ad text paraphrasing, containing human preference data, to enable the analysis of the linguistic factors and to support the development of methods for generating attractive ad texts. Compared with v1.0, this dataset is 20 times larger, comprising 16,460 ad text paraphrase pairs, each annotated with preference data from ten evaluators, thereby enabling a more comprehensive and reliable analysis. Through the experiments, we identified multiple linguistic features of engaging ad texts that were not observed in v1.0 and explored various methods for generating attractive ad texts. Furthermore, our analysis demonstrated the relationships between human preference and ad performance, and highlighted the potential of reference-free metrics based on large language models for evaluating ad text attractiveness. The dataset is publicly available at: https://github.com/CyberAgentAILab/AdParaphrase-v2.0.

PredictaBoard: Benchmarking LLM Score Predictability

Lorenzo Pacchiardi, Konstantinos Voudouris, Ben Slater, Fernando Martínez-Plumed, Jose Hernandez-Orallo, Lexin Zhou, Wout Schellaert

Despite possessing impressive skills, Large Language Models (LLMs) often fail unpre-dictably, demonstrating inconsistent success in even basic common sense reasoning tasks. This unpredictability poses a significant challenge to ensuring their safe deployment, as identifying and operating within a reliable “safe zone” is essential for mitigating risks. To address this, we present PredictaBoard, a novel collabo-rative benchmarking framework designed to evaluate the ability of score predictors (referred to as assessors) to anticipate LLM errors on specific task instances (i.e., prompts) from existing datasets. PredictaBoard evaluates pairs of LLMs and assessors by considering the rejection rate at different tolerance errors. As such, PredictaBoard stimulates research into developing better assessors and making LLMs more predictable, not only with a higher average performance. We conduct illustrative experiments using baseline assessors and state-of-the-art LLMs. PredictaBoard highlights the critical need to evaluate predictability alongside performance, paving the way for safer AI systems where errors are not only minimised but also anticipated and effectively mitigated. Code for our bench-mark can be found at https://github. com/Kinds-of-Intelligence-CFI/PredictaBoard

How Far are LLMs from Being Our Digital Twins? A Benchmark for Persona-Based Behavior Chain Simulation

Rui Li, Heming Xia, Xinfeng Yuan, Qingxiu Dong, Lei Sha, Wenjie Li, Zhifang Sui

Recently, LLMs have garnered increasing attention across academic disciplines for their potential as human digital twins, virtual proxies designed to replicate individuals and autonomously perform tasks such as decision-making, problem-solving, and reasoning on their behalf. However, current evaluations of LLMs primarily emphasize dialogue simulation while overlooking human behavior simulation, which is crucial for digital twins. To address this gap, we introduce BehaviorChain, the first benchmark for evaluating LLMs' ability to simulate continuous human behavior. BehaviorChain comprises diverse, high-quality, persona-based behavior chains, totaling 15,846 distinct behaviors across 1,001 unique personas, each with detailed history and profile metadata. For evaluation, we integrate persona metadata into LLMs and employ them to iteratively infer contextually appropriate behaviors within dynamic scenarios provided by BehaviorChain. Comprehensive evaluation results demonstrated that even state-of-the-art models struggle with accurately simulating continuous human behavior.

BOSE: A Systematic Evaluation Method Optimized for Base Models

Hongzhi Luan, Changxin Tian, Zhaoxin Huan, Xiaolu Zhang, Kunlong Chen, Zhiqiang Zhang, JUN ZHOU

This paper poses two critical issues in evaluating base models (without post-training): (1) Unstable evaluation during training: in the early stages of pre-training, the models lack the capability to answer questions as required, leading to unstable evaluation results. This instability makes it difficult to provide solid conclusions to guide the training, especially for key experiments such as data ablation and scaling law. (2) Inconsistency between base and instruct models: base models generally exhibit poorer evaluation performance compared to corresponding instruct models. This gap poses a challenge for assessing whether a base model with better evaluation can truly lead to a better instruct model. To address these issues, we propose **B**ase model **O**riented **S**ystematic **E**valuation (**BOSE**), a method specifically designed to optimize the evaluation of base models. Specifically, BOSE introduces two key innovations: In-Context Light-instruction Prompt (**ICLiP**) for open-ended tasks and **Blank-ppl** for multi-choice tasks with candidate options, which transforms the standard perplexity (ppl) metric into a fill-in-the-blank format to mitigate early-stage evaluation fluctuations. Furthermore, we are the first to propose Kendall's rank correlation to quantitatively measure the evaluation stability and consistency. Experimental results demonstrate that BOSE significantly enhances both the stability of evaluations during pre-training and the consistency between base and instruct models, thereby providing more reliable guidance for the LLMs' training.

A Large and Balanced Corpus for Fine-grained Arabic Readability Assessment

Khalid Elmadani, Nizar Habash, Hanada Taha

This paper introduces the Balanced Arabic Readability Evaluation Corpus BAREC, a large-scale, fine-grained dataset for Arabic readability assessment. BAREC consists of 68,182 sentences spanning 1+ million words, carefully curated to cover 19 readability levels, from kindergarten to postgraduate comprehension. The corpus balances genre diversity, topical coverage, and target audiences, offering a comprehensive resource for evaluating Arabic text complexity. The corpus was fully manually annotated by a large team of annotators. The average pairwise inter-annotator agreement, measured by Quadratic Weighted Kappa, is 81.3%, reflecting a high level of substantial agreement. Beyond presenting the corpus, we benchmark automatic readability assessment across different granularity levels, comparing a range of techniques. Our results highlight the challenges and opportunities in Arabic readability modeling, demonstrating competitive performance across various methods. To support research and education, we will make BAREC openly available, along with detailed annotation guidelines and benchmark results.

Evaluating LLMs' Assessment of Mixed-Context Hallucination Through the Lens of Summarization

Siya Qi, RUI CAO, Yulan He, Zheng Yuan

With the rapid development of large language models (LLMs), LLM-as-a-judge has emerged as a widely adopted approach for text quality evaluation, including hallucination evaluation. While previous studies have focused exclusively on single-context evaluation (e.g., discourse faithfulness or world factuality), real-world hallucinations typically involve mixed contexts, which remains inadequately evaluated. In this study, we use summarization as a representative task to comprehensively evaluate LLMs' capability in detecting mixed-context hallucinations, specifically distinguishing between factual and non-factual hallucinations. Through extensive experiments across direct generation and retrieval-based models of varying scales, our main observations are: (1) LLMs' intrinsic knowledge introduces inherent biases in hallucination evaluation; (2) These biases particularly impact the detection of factual hallucinations, yielding a significant performance bottleneck; (3) The fundamental challenge lies in effective knowledge utilization, balancing between LLMs' intrinsic knowledge and external context for accurate mixed-context hallucination evaluation.

Review-Instruct: A Review-Driven Multi-Turn Conversations Generation Method for Large Language Models

Jiangxu Wu, Cong Wang, Tianhuang Su, lin haozhi, JunYang, Zhangchao, Binqiang Pan, SongpanYang, Mingpeng, Kai Shi, ZIXIAN LI

The effectiveness of large language models (LLMs) in conversational AI is hindered by their reliance on single-turn supervised fine-tuning (SFT) data, which limits contextual coherence in multi-turn dialogues. Existing methods for generating multi-turn dialogue data struggle to ensure both diversity and quality in instructions. To address this, we propose Review-Instruct, a novel framework that synthesizes multi-turn conversations through an iterative "Ask-Respond-Review" process involving three agent roles: a Candidate, multiple Reviewers, and a Chairman. The framework iteratively refines instructions by incorporating Reviewer feedback, enhancing dialogue diversity and difficulty. We construct a multi-turn dataset using the Alpaca dataset and fine-tune the LLaMA2-13B model. Evaluations on MT-Bench, MMLU-Pro, and Auto-Arena demonstrate significant improvements, achieving absolute gains of 2.9\% on MMLU-Pro and 2\% on MT-Bench compared to prior state-of-the-art models based on LLaMA2-13B. Ablation studies confirm the critical role of the Review stage and the use of multiple Reviewers in boosting instruction diversity and difficulty. Our work highlights the potential of review-driven, multi-agent frameworks for generating high-quality conversational data at scale.

EuroVerdict: A Multilingual Dataset for Verdict Generation Against Misinformation

Daniel Russo, Fariba Sadeghi, Stefano Menini, Marco Guerini

Misinformation is a global issue that shapes public discourse, influencing opinions and decision-making across various domains. While automated fact-checking (AFC) has become essential in combating misinformation, most work in multilingual settings has focused on claim verification rather than generating explanatory verdicts, leaving a gap in AFC resources beyond English. To address this, we introduce EuroVerdict, a multilingual dataset designed for verdict generation, covering eight European languages. Developed in collaboration with professional fact-checkers, the dataset includes claims, manually written verdicts, and supporting evidence, consisting of fact-checking articles and additional secondary sources. We evaluate EuroVerdict with Llama-3.1-8B-Instruct on verdict generation under different settings, varying the prompt language, input article language, and training approach. Our results show that fine-tuning consistently improves performance, with models fine-tuned on original-language articles achieving the highest scores in both automatic and human evaluations. Using articles in a different language than the claim slightly lowers performance; however, pairing them with language-specific prompts improves results. Zero-shot and Chain of Thought configurations perform worse, reinforcing the benefits of fine-tuning for multilingual verdict generation.

LoFTI: Localization and Factuality Transfer to Indian Locales

Sona Elza Simon, Soumen Kumar Mondal, Abhishek Singhania, Sayambhu Sen, Preethi Jyothi

Large language models (LLMs) encode vast amounts of world knowledge acquired via training on large web-scale datasets crawled from the internet. However, the datasets used to train the LLMs typically exhibit a geographical bias towards English-speaking Western countries. This results in LLMs producing biased or hallucinated responses to queries that require answers localized to other geographical regions. In this work, we introduce a new benchmark named LoFTI (Localization and Factuality Transfer to Indian Locales) that can be used to evaluate an LLM's contextual localization and factual text transfer capabilities. LoFTI consists of factual statements about entities in source and target locations; the source locations are spread across the globe and the target locations are all within India with varying degrees of hyperlocality (country, states, cities). The entities span a wide variety of categories. We use LoFTI to evaluate Mixtral, Llama3.3-70B, GPT-4 and two other Mixtral-based approaches well-suited to the task of localized factual transfer. We demonstrate that LoFTI is a high-quality evaluation benchmark and all the models, including GPT-4, produce skewed results across varying levels of hyperlocality.

Improving Meta Introspection of Small LLMs by Learning Self-Reflection from Self-Generated Data

Jiaqi Li, Xinyi Dong, Yang Liu, Zhizhuo Yang, Quansen Wang, Xiaobo Wang, Song-Chun Zhu, Zixia Jia, Zilong Zheng

We present a novel pipeline, ReflectEvo, to demonstrate that small language models (SLMs) can enhance meta introspection through reflection learning. This process iteratively generates self-reflection for self-training, fostering a continuous and self-evolving process. Leveraging this pipeline, we construct ReflectEvo-460k, a large-scale, comprehensive, self-generated reflection dataset with broadened instructions and diverse multi-domain tasks. Building upon this dataset, we demonstrate the effectiveness of reflection learning to improve SLMs' reasoning abilities using SFT and DPO with remarkable performance, substantially boosting Llama-3 from 52.4% to 71.2% and Mistral from 44.4% to 71.1%. It validates that ReflectEvo can rival or even surpass the reasoning capability of the three prominent open-sourced models on BIG-bench without distillation from superior models or fine-grained human annotation. We further conduct a deeper analysis of the high quality of self-generated reflections and their impact on error localization and correction. Our work highlights the potential of continuously enhancing the reasoning performance of SLMs through iterative reflection learning in the long run.

GaRAGe: A Benchmark with Grounding Annotations for RAG Evaluation

Ionut Teodor Sorodoc, Leonardo F. R. Ribeiro, Rexhina Blloshmi, Christopher Davis, Adrià de Gispert

We present GaRAGe, a large RAG benchmark with human-curated long-form answers and annotations of each grounding passage, allowing a fine-grained evaluation of whether LLMs can identify relevant grounding when generating RAG answers. Our benchmark contains 2366 questions of diverse complexity, dynamism, and topics, and includes over 35K annotated passages retrieved from both private document sets and the Web, to reflect real-world RAG use cases. This makes it an ideal test bed to evaluate an LLM's ability to identify only the relevant information necessary to compose a response, or provide a deflective response when there is insufficient information. Evaluations of multiple state-of-the-art LLMs on GaRAGe show that the models tend to over-summarise rather than (a) ground their answers strictly on the annotated relevant passages (reaching at most a Relevance-Aware Factuality Score of 60%), or (b) deflect when no relevant grounding is available (reaching at most 31% true positive rate in deflections). The F_{1} in attribution to relevant sources is at most 58.9%, and we show that performance is particularly reduced when answering time-sensitive questions and when having to draw knowledge from sparser private grounding sources.

RAG-RewardBench: Benchmarking Reward Models in Retrieval Augmented Generation for Preference Alignment

Zhuoran Jin, Hongbang Yuan, Tianyi Men, Pengfei Cao, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, Jun Zhao

Despite the significant progress made by existing retrieval augmented language models (RALMs) in providing trustworthy responses and grounding in reliable sources, they often overlook effective alignment with human preferences. In the alignment process, reward models (RMs) act as a crucial proxy for human values to guide optimization. However, it remains unclear how to evaluate and select a reliable RM for preference alignment in RALMs. To this end, we propose RAG-RewardBench, the first benchmark for evaluating RMs in RAG settings. First, we design four crucial and challenging RAG-specific scenarios to assess RMs, including multi-hop reasoning, fine-grained citation, appropriate abstain, and conflict robustness. Then, we incorporate 18 RAG subsets, six retrievers, and 24 RALMs to increase the diversity of data sources. Finally, we adopt an LLM-as-a-judge approach to improve preference annotation efficiency and effectiveness, exhibiting a strong correlation with human annotations. Based on the RAG-RewardBench, we conduct a comprehensive evaluation of 45 RMs and uncover their limitations in RAG scenarios. Additionally, we also reveal that existing trained RALMs show almost no improvement in preference alignment, highlighting the need for a shift towards preference-aligned training.

PAM: Paraphrase AMR-Centric Evaluation Metric

Afonso Sousa, Henrique Lopes Cardoso

Paraphrasing is rooted in semantics, which makes evaluating paraphrase generation systems hard. Current paraphrase generators are typically evaluated using borrowed metrics from adjacent text-to-text tasks, like machine translation or text summarization. These metrics tend to have ties to the surface form of the reference text. This is not ideal for paraphrases as we typically want variation in the lexicon while persisting semantics. To address this problem, and inspired by learned similarity evaluation on plain text, we propose PAM, a Paraphrase AMR-Centric Evaluation Metric. This metric uses AMR graphs extracted from the input text, which consist of semantic structures agnostic to the text surface form, making the resulting evaluation metric more robust to variations in syntax or lexicon. Additionally, we evaluated PAM on different semantic textual similarity datasets and found that it improves the correlations with human semantic scores when compared to other AMR-based metrics.

Filling the Temporal Void: Recovering Missing Publication Years in the Project Gutenberg Corpus Using LLMs

Omar Momen, Manuel Schaaf, Alexander Mehler

Analyzing texts spanning long periods of time is critical for researchers in historical linguistics and related disciplines. However, publicly available corpora suitable for such analyses are scarce. The Project Gutenberg (PG) corpus presents a significant yet underutilized opportunity in this context, due to the absence of accurate temporal metadata. We take advantage of language models and information retrieval to explore four sources of information -- Open Web, Wikipedia, Open Library API, and Project Gutenberg (PG) -- to add missing temporal metadata to PG books. Through 20 experiments employing state-of-the-art Large Language Models (LLMs) and Retrieval Augmented Generation (RAG) methods, we estimate the production years of all PG books. We curate an enriched metadata repository for PG books and propose a refined version of the PG corpus, which includes 53774 books with a total of 3.8 billion tokens in 11 languages, produced between 1600 and 2000. This work provides a new resource for computational linguistics and humanities studies focusing on diachronic analyses. All experiments data are published for the public (https://drive.proton.me/urls/DNQE5562RM#L2ldqrQ88Y6B).

ExpliCa: Evaluating Explicit Causal Reasoning in Large Language Models

Martina Miliani, Serena Auriemma, Alessandro Bondielli, Emmanuele Chersoni, Lucia Passaro, Irene Sucameli, Alessandro Lenci

Large Language Models (LLMs) are increasingly used in tasks requiring interpretive and inferential accuracy. In this paper, we introduce ExpliCa, a new dataset for evaluating LLMs in explicit causal reasoning. ExpliCa uniquely integrates both causal and temporal relations presented in different linguistic orders and explicitly expressed by linguistic connectives. The dataset is enriched with crowdsourced human acceptability ratings. We tested LLMs on ExpliCa through prompting and perplexity-based metrics. We assessed seven commercial and open-source LLMs, revealing that even top models struggle to reach 0.80 accuracy. Interestingly, models tend to confound temporal relations with causal ones, and their performance is also strongly influenced by the linguistic order of the events. Finally, perplexity-based scores and prompting performance are differently affected by model size.

MC-MKE: A Fine-Grained Multimodal Knowledge Editing Benchmark Emphasizing Modality Consistency

Junzhe Zhang, Huixuan Zhang, Xunjian Yin, Baizhou Huang, Xu Zhang, Xinyu Hu, Xiaojun Wan

Multimodal large language models (MLLMs) are prone to non-factual or outdated knowledge issues, highlighting the importance of knowledge editing. Many benchmark has been proposed for researching multimodal knowledge editing. However, previous benchmarks focus on limited scenarios due to the lack of rigorous definition of multimodal knowledge. To better evaluate multimodal knowledge editing, we propose a decomposed definition of multimodal knowledge. Following the decomposed definition of multimodal knowledge, we introduce three scenarios and a novel requirement modality consistency. We construct MC-MKE, a fine-grained **M**ultimodal **K**nowledge **E**diting benchmark emphasizing **M**odality **C**onsistency through strict data selection. We evaluate four multimodal knowledge editing methods on MC-MKE, revealing their limitations, particularly in terms of modality consistency. Our work highlights the challenges posed by multimodal knowledge editing and motivates further research in developing effective techniques for this task.

Are Your LLMs Capable of Stable Reasoning?

Junnan Liu, Hongwei Liu, Linchen Xiao, Ziyi Wang, Kuikun Liu, Songyang Gao, Wenwei Zhang, Songyang Zhang, Kai Chen

The rapid advancement of large language models (LLMs) has shown remarkable progress in complex reasoning tasks. However, a significant disparity exists between benchmark performances and real-world applications. We attribute this gap primarily to current evaluation protocols and metrics, which inadequately capture the full spectrum of LLM capabilities, especially in complex reasoning tasks where both accuracy and consistency are essential. In this paper, we introduce **G-Pass@**k, a novel evaluation metric that continuously assesses model performance across multiple sampling attempts, quantifying both the model's performance potential and its stability. Through extensive experiments on various public and newly constructed benchmarks, we employ G-Pass@k in conjunction with state-of-the-art large language models to provide comprehensive insights into their potential capabilities and operational consistency. Our findings reveal a significant opportunity to enhance the realistic reasoning abilities of LLMs, underscoring the necessity for more robust evaluation metrics.

Annotating the Annotators: Analysis, Insights and Modelling from an Annotation Campaign on Persuasion Techniques Detection

Davide Bassi, Dimitar Iliyanov Dimitrov, Bernardo D'Auria, Firoj Alam, Maram Hasanain, Christian Moro, Luisa Orrù, Gian Piero Turchi, Preslav Nakov, Giovanni Da San Martino

Persuasion (or propaganda) techniques detection is a relatively novel task in Natural Language Processing (NLP). While there have already been a number of annotation campaigns, they have been based on heuristic guidelines, which have never been thoroughly discussed. Here, we present the first systematic analysis of a complex annotation task -detecting 22 persuasion techniques in memes-, for which we provided continuous expert oversight. The presence of an expert allowed us to critically analyze specific aspects of the annotation process. Among our findings, we show that inter-annotator agreement alone inadequately assessed annotation correctness. We thus define and track different error types, revealing that expert feedback shows varying effectiveness across error categories. This pattern suggests that distinct mechanisms underlie different kinds of misannotations. Based on our findings, we advocate for the necessity of expert oversight in annotation tasks and periodic quality audits. In order to reduce the costs for this, we introduce a probabilistic model for optimizing intervention scheduling. To support future research on annotation quality's impact on model performance, we release our comprehensive dataset containing all individual annotator labels, all consolidated annotations, and all expert revisions.

Benchmarking the Benchmarks: Reproducing Climate-Related NLP Tasks

Tom Calamai, Oana Balalau, Fabian M. Suchanek

Nowadays, news, legislation, guidelines, and corporate reports are discussing the impact of climate change, how to best prevent future damage and adapt to a changing environment. Navigating this wealth of information can be overwhelming for citizens, journalists, legislators, and other stakeholders. Hence, significant efforts have been made in the natural language processing community to facilitate the automatic analysis of a large corpus of text via tasks such as climate-related topic detection, climate risk classification, question answering over climate topics, and many more. In this work, we perform a reproducibility study on 8 tasks and 29 datasets, testing 6 models. Our goal is to highlight potential issues within datasets and evaluation methodologies while also providing cleaned versions of those datasets and a reliable testing platform. We identify in each dataset annotation errors, which can be leveraged to provide new and higher quality datasets. Finally, we release a toolbox containing the cleaned datasets and the evaluation code of models, together with a pipeline for applying the models on corporate documents.

PersonaLens: A Benchmark for Personalization Evaluation in Conversational AI Assistants

Zheng Zhao, Clara Vania, Subhradeep Kayal, Naila Khan, Shay B Cohen, Emine Yilmaz

Large language models (LLMs) have advanced conversational AI assistants. However, systematically evaluating how well these assistants apply personalization—adapting to individual user preferences while completing tasks—remains challenging. Existing personalization benchmarks focus on chit-chat, non-conversational tasks, or narrow domains, failing to capture the complexities of personalized task-oriented assistance. To address this, we introduce PersonaLens, a comprehensive benchmark for evaluating personalization in task-oriented AI assistants. Our benchmark features diverse user profiles equipped with rich preferences and interaction histories, along with two specialized LLM-based agents: a user agent that engages in realistic task-oriented dialogues with AI assistants, and a judge agent that employs the LLM-as-a-Judge paradigm to assess personalization, response quality, and task success. Through extensive experiments with current LLM assistants across diverse tasks, we reveal significant variability in their personalization capabilities, providing crucial insights for advancing conversational AI systems.

Call for Rigor in Reporting Quality of Instruction Tuning Data

Hyeonseok Moon, Jaehyung Seo, Heuiseok Lim

Instruction tuning is crucial for adapting large language models (LLMs) to align with user intentions. Numerous studies emphasize the significance of the quality of instruction tuning (IT) data, revealing a strong correlation between IT data quality and the alignment performance of LLMs. In these studies, the quality of IT data is typically assessed by evaluating the performance of LLMs trained with that data. However, we identified a prevalent issue in such practice: hyperparameters for training models are often selected arbitrarily without adequate justification. We observed significant variations in hyperparameters applied across different studies, even when training the same model with the same data. In this study, we demonstrate the potential problems arising from this practice and emphasize the need for careful consideration in verifying data quality. Through our experiments on the quality of LIMA data and a selected set of 1,000 Alpaca data points, we demonstrate that arbitrary hyperparameter decisions can make any arbitrary conclusion.

Crowd Comparative Reasoning: Unlocking Comprehensive Evaluations for LLM-as-a-Judge

Qiyuan Zhang, Yufei Wang, Yuxin Jiang, Liangyou Li, Chuhan Wu, Yasheng Wang, Xin Jiang, Lifeng Shang, Ruiming Tang, Fuyuan Lyu, Chen Ma

LLM-as-a-Judge, which generates chain-of-thought (CoT) judgments, has become a widely adopted auto-evaluation method. However, its reliability is compromised by the CoT reasoning's inability to capture comprehensive and deeper details, often leading to incomplete outcomes. Existing methods mainly rely on majority voting or criteria expansion, which is insufficient to address the limitation in CoT. We propose Crowd-based Comparative Evaluation, which introduces additional crowd responses to compare with the candidate responses, thereby exposing deeper and more comprehensive details within the candidate responses. This process effectively guides LLM-as-a-Judge to provide a more detailed CoT judgment. Extensive experiments demonstrate that our approach enhances evaluation reliability, achieving an average accuracy gain of 6.7% across five benchmarks. Moreover, our method produces higher-quality CoTs that facilitate judge distillation and exhibit superior performance in rejection sampling for supervised fine-tuning (SFT), referred to as crowd rejection sampling, thereby enabling more efficient SFT. Our analysis confirms that CoTs generated by ours are more comprehensive and of higher quality, and evaluation accuracy improves as inference scales.

L-CiteEval: A Suite for Evaluating Fidelity of Long-context Models

Zecheng Tang, Keyan Zhou, Juntao Li, Baibei Ji, jianye hou, Min Zhang

Long-context models(LCMs) have witnessed remarkable advancements in recent years, facilitating real-world tasks like long-document QA. The success of LCMs is founded on the hypothesis that the model demonstrates strong \textbf{fidelity}, enabling it to respond based on the provided long context rather than relying solely on the intrinsic knowledge acquired during pre-training. Yet, in this paper, we find that open-sourced LCMs are not as faithful as expected. We introduce \textit{L-CiteEval}, an out-of-the-box suite that can assess both generation quality and fidelity in long-context understanding tasks. It covers 11 tasks with context lengths ranging from 8K to 48K and a corresponding automatic evaluation pipeline. Evaluation of 11 cutting-edge closed-source and open-source LCMs indicates that, while there are minor differences in their generation, open-source models significantly lag behind closed-source counterparts in terms of fidelity. Furthermore, we analyze the benefits of citation generation for LCMs from both the perspective of explicit model output and the internal attention mechanism.

HoH: A Dynamic Benchmark for Evaluating the Impact of Outdated Information on Retrieval-Augmented Generation

Jie Ouyang, Tingyue Pan, Mingyue Cheng, Ruiran Yan, Yucong Luo, Jiaying Lin, Qi Liu

While Retrieval-Augmented Generation (RAG) has emerged as an effective approach for addressing the knowledge outdating problem in Large Language Models (LLMs), it still faces a critical challenge: the prevalence of outdated information in knowledge bases. Current research primarily focuses on incorporating up-to-date information, yet the impact of outdated information coexisting in retrieval sources remains inadequately addressed. To bridge this gap, we introduce HoH, the first benchmark specifically designed to evaluate the impact of outdated information on RAG. Our benchmark leverages token-level diff algorithms combined with LLM pipelines to efficiently create a large-scale QA dataset that accurately captures the evolution of temporal knowledge in real-world facts. Through comprehensive experiments, we reveal that outdated information significantly degrades RAG performance in two critical ways: (1) it substantially reduces response accuracy by distracting models from correct information, and (2) it can mislead models into generating potentially harmful outputs, even when current information is available. Current RAG approaches struggle with both retrieval and generation aspects when handling outdated information. These findings highlight the urgent need for innovative solutions to address the temporal challenges in RAG.

AutoMedEval: Harnessing Language Models for Automatic Medical Capability Evaluation

Xiechi Zhang, Zetian Ouyang, Linlin Wang, Gerard de Melo, Zhu Cao, Xiaoling Wang, Ya Zhang, Yanfeng Wang, Liang He

With the proliferation of large language models (LLMs) in the medical domain, there is increasing demand for improved evaluation techniques to assess their capabilities. However, traditional metrics like F1 and ROUGE, which rely on token overlaps to measure quality, significantly overlook the importance of medical terminology. While human evaluation tends to be more reliable, it can be very costly and may as well suffer from inaccuracies due to limits in human expertise and motivation. Although there are some evaluation methods based on LLMs, their usability in the medical field is limited due to their proprietary nature or lack of expertise. To tackle these challenges, we present AutoMedEval, an open-sourced automatic evaluation model with 13B parameters specifically engineered to measure the question-answering proficiency of medical LLMs. The overarching objective of AutoMedEval is to assess the quality of responses produced by diverse models, aspiring to significantly reduce the dependence on human evaluation. Specifically, we propose a hierarchical training method involving curriculum instruction tuning and an iterative knowledge introspection mechanism, enabling AutoMedEval to acquire professional medical assessment capabilities with limited instructional data. Human evaluations indicate that AutoMedEval surpasses other baselines in terms of correlation with human judgments.

Evaluating Multimodal Large Language Models on Video Captioning via Monte Carlo Tree Search

Linhao Yu, Xingguang Ji, Yahui Liu, Fanheng Kong, Chenxi Sun, Jingyuan Zhang, Hongzhi Zhang, V. W., Fuzheng Zhang, Deyi Xiong

Video captioning can be used to assess the video understanding capabilities of Multimodal Large Language Models (MLLMs). However, existing benchmarks and evaluation protocols suffer from crucial issues, such as inadequate or homogeneous creation of key points, exorbitant cost of data creation, and limited evaluation scopes. To address these issues, we propose an automatic framework, named AutoCaption, which leverages Monte Carlo Tree Search (MCTS) to construct numerous and diverse descriptive sentences (\textit{i.e.}, key points) that thoroughly represent video content in an iterative way. This iterative captioning strategy enables the continuous enhancement of video details such as actions, objects' attributes, environment details, etc. We apply AutoCaption to curate MCTS-VCB, a fine-grained video caption benchmark covering video details, thereby enabling a comprehensive evaluation of MLLMs on the video captioning task. We evaluate more than 20 open- and closed-source MLLMs of varying sizes on MCTS-VCB. Results show that MCTS-VCB can effectively and comprehensively evaluate the video captioning capability, with Gemini-1.5-Pro achieving the highest F1 score of 71.2. Interestingly, we fine-tune InternVL2.5-8B with the AutoCaption-generated data, which helps the model achieve an overall improvement of 25.0% on MCTS-VCB and 16.3% on DREAM-1K, further demonstrating the effectiveness of AutoCaption. The code and data are available at .

AlignMMBench: Evaluating Chinese Multimodal Alignment in Large Vision-Language Models

Yuhang Wu, Wenmeng Yu, Yean Cheng, Yan Wang, Xiaohan Zhang, Jiazheng Xu, Ming Ding, Yuxiao Dong

Evaluating the alignment capabilities of large Vision-Language Models (VLMs) is essential for determining their effectiveness as helpful assistants. However, existing benchmarks primarily focus on basic abilities using nonverbal methods, such as yes-no and multiple-choice questions. In this paper, we address this gap by introducing AlignMMBench, which provides more nuanced evaluations of alignment capabilities and is the first benchmark specifically designed for Chinese visual contexts. This benchmark is meticulously curated from real-world scenarios and internet sources, encompassing thirteen specific tasks across three categories, and includes both single-turn and multi-turn dialogue scenarios. Incorporating a prompt rewrite strategy, AlignMMBench encompasses 1,054 images and 4,978 question-answer pairs. To facilitate the evaluation pipeline, we develop CritiqueVLM, a rule-calibrated evaluator that exceeds GPT-4's evaluation ability. Additionally, we measure the “alignment score”, a quantitative metric designed to assess the robustness and stability of models across diverse prompts. Finally, we evaluate the performance of representative VLMs on AlignMMBench, offering insights into the capabilities and limitations of different VLM architectures. The evaluation code and data are available at https://github.com/THUDM/AlignMMBench.

That is Unacceptable: the Moral Foundations of Canceling

Soda Marem Lo, Oscar Araque, Rajesh Sharma, Marco Antonio Stranisci

Canceling is a morally-driven phenomenon that hinders the development of safe social media platforms and contributes to ideological polarization. To address this issue we present the Canceling Attitudes Detection (CADE) dataset, an annotated corpus of canceling incidents aimed at exploring the factors of disagreements in evaluating people's canceling attitudes on social media. Specifically, we study the impact of annotators' morality in their perception of canceling, showing that morality is an independent axis for the explanation of disagreement on this phenomenon. Annotator's judgments heavily depend on the type of controversial events and involved celebrities. This shows the need to develop more event-centric datasets to better understand how harms are perpetrated in social media and to develop more aware technologies for their detection.

Identifying Reliable Evaluation Metrics for Scientific Text Revision

Leane Jourdan, Nicolas Hernandez, Florian Boudin, Richard Dufour

Evaluating text revision in scientific writing remains a challenge, as traditional metrics such as ROUGE and BERTScore primarily focus on similarity rather than capturing meaningful improvements. In this work, we analyse and identify the limitations of these metrics and explore alternative evaluation methods that better align with human judgments. We first conduct a manual annotation study to assess the quality of different revisions. Then, we investigate reference-free evaluation metrics from related NLP domains. Additionally, we examine LLM-as-a-judge approaches, analysing their ability to assess revisions with and without a gold reference. Our results show that LLMs effectively assess instruction-following but struggle with correctness, while domain-specific metrics provide complementary insights. We find that a hybrid approach combining LLM-as-a-judge evaluation and task-specific metrics offers the most reliable assessment of revision.

Hybrid Preferences: Learning to Route Instances for Human vs. AI Feedback

Lester James Validad Miranda, Yizhong Wang, Yanai Elazar, Sachin Kumar, Valentina Pyatkin, Faeze Brahman, Noah A. Smith, Hannaneh Hajishirzi, Pradeep Dasigi

Learning from human feedback has enabled the alignment of language models (LMs) with human preferences. However, collecting human preferences is expensive and time-consuming, with highly variable annotation quality. An appealing alternative is to distill preferences from LMs as a source of synthetic annotations, offering a cost-effective and scalable alternative, albeit susceptible to other biases and errors. In this work, we introduce HyPER, a Hybrid Preference routER that defers an annotation to either humans or LMs, achieving better annotation quality while reducing the cost of human-only annotation. We formulate this as an optimization problem: given a preference dataset and an evaluation metric, we (1) train a performance prediction model (PPM) to predict a reward model's (RM) performance on an arbitrary combination of human and LM annotations and (2) employ a routing strategy that selects a combination that maximizes predicted performance. We train the PPM on MultiPref, a new preference dataset with 10K instances paired with human and LM labels. We show that the selected hybrid mixture of synthetic and direct human preferences using HyPER achieves better RM performance compared to using either one exclusively by 7-13% on RewardBench and generalizes across unseen preference datasets and other base models. We also observe the same trend in other benchmarks using Best-of-N reranking, where the hybrid mix has 2-3% better performance. Finally, we analyze features from HyPER and find that prompts with moderate safety concerns or complexity benefit the most from human feedback.

The UD-NewsCrawl Treebank: Reflections and Challenges from a Large-scale Tagalog Syntactic Annotation Project

Angelina Aspra Aquino, Lester James Validad Miranda, Elsie Marie T. Or

This paper presents UD-NewsCrawl, the largest Tagalog treebank to date, containing 15.6k trees manually annotated according to the Universal Dependencies framework. We detail our treebank development process, including data collection, pre-processing, manual annotation, and quality assurance procedures. We provide baseline evaluations using multiple transformer-based models to assess the performance of state-of-the-art dependency parsers on Tagalog. We also highlight challenges in the syntactic analysis of Tagalog given its distinctive grammatical properties, and discuss its implications for the annotation of this treebank. We anticipate that UD-NewsCrawl and our baseline model implementations will serve as valuable resources for advancing computational linguistics research in underrepresented languages like Tagalog.

Beyond Outcomes: Transparent Assessment of LLM Reasoning in Games

Wenye Lin, Jonathan Roberts, Yunhan Yang, Samuel Albanie, Zongqing Lu, Kai Han

Large Language Models (LLMs) are increasingly deployed in real-world applications that demand complex reasoning. To track progress, robust benchmarks are required to evaluate their capabilities beyond superficial pattern recognition. However, current LLM reasoning benchmarks often face challenges such as insufficient interpretability, performance saturation or data contamination. To address these challenges, we introduce GAMEBoT, a gaming arena designed for rigorous and transparent assessment of LLM reasoning capabilities. GAMEBoT decompose complex reasoning in games into predefined modular subproblems. This decomposition allows us to design a suite of Chain-of-Thought (CoT) prompts infused with domain knowledge to guide LLMs in addressing these subproblems before action selection. Furthermore, we develop a suite of rule-based algorithms to generate ground truth for these subproblems, enabling rigorous validation of the LLMs' intermediate reasoning steps. This approach facilitates evaluation of both the quality of final actions and the accuracy of the underlying reasoning process. GAMEBoT also naturally alleviates the risk of data contamination through dynamic games and head-to-head LLM competitions. We benchmark 17 prominent LLMs across eight games, encompassing various strategic abilities and game characteristics. Our results suggest that GAMEBoT presents a significant challenge, even when LLMs are provided with detailed CoT prompts.

Commonsense Reasoning in Arab Culture

Abdelrahman Sadallah, Junior Cedric Tonga, Khalid Almubarak, Saeed Almheiri, Farah Atif, Chatrine Qwaider, Karima Kadaoui, Sara Shatnawi, Yaser Alesh, Fajri Koto

Despite progress in Arabic large language models, such as Jais and AceGPT, their evaluation on commonsense reasoning has largely relied on machine-translated datasets, which lack cultural depth and may introduce Anglocentric biases. Commonsense reasoning is shaped by geographical and cultural contexts, and existing English datasets fail to capture the diversity of the Arab world. To address this, we introduce \datasetname, a commonsense reasoning dataset in Modern Standard Arabic (MSA), covering cultures of 13 countries across the Gulf, Levant, North Africa, and the Nile Valley. The dataset was built from scratch by engaging native speakers to write and validate culturally relevant questions for their respective countries. \datasetname spans 12 daily life domains with 54 fine-grained subtopics, reflecting various aspects of social norms, traditions, and everyday experiences. Zero-shot evaluations show that open-weight language models with up to 32B parameters struggle to comprehend diverse Arab cultures, with performance varying across regions. These findings highlight the need for more culturally aware models and datasets tailored to the Arabic-speaking world.

Building a Long Text Privacy Policy Corpus with Multi-Class Labels

David Stein, Florencia Marotta-Wurgler

Legal text poses distinctive challenges for natural language processing. The legal import of a term may depend on omissions, cross-references, or silence, Further, legal text is often susceptible to multiple valid, conflicting interpretations; as the saying goes: a good lawyer's answer to any question is "it depends." This work introduces a new, hand-coded dataset for the interpretation of privacy policies. It includes privacy policies from 149 firms, including materials incorporated by reference. The policies are annotated across 64 dimension that reflect the applicable legal rules and contested terms from EU and US privacy regulation and litigation. Our annotation methodology is designed to capture the capture core challenges peculiar to legal language, including indeterminacy, interdependence between clauses, meaningful silence, and the implications of legal defaults. We present a set of baseline results for the dataset using current large language models.

Browsing Lost Unformed Recollections: A Benchmark for Tip-of-the-Tongue Search and Reasoning

Sky CH-Wang, Darshan Girish Deshpande, Smaranda Muresan, Anand Kannappan, Rebecca Qian

We introduce Browsing Lost Unformed Recollections, a tip-of-the-tongue known-item search and reasoning benchmark for general AI assistants. BLUR introduces a set of 573 real-world validated questions that demand searching and reasoning across multimodal and multilingual inputs, as well as proficient tool use, in order to excel on. Humans easily ace these questions (scoring on average 98%), while the best-performing system scores around 56%. To facilitate progress toward addressing this challenging and aspirational use case for general AI assistants, we release 350 questions through a public leaderboard, retain the answers to 250 of them, and have the rest as a private test set.

LLMs can Perform Multi-Dimensional Analytic Writing Assessments: A Case Study of L2 Graduate-Level Academic English Writing

Zhengxiang Wang, Veronika Makarova, Zhi Li, Jordan Kodner, Owen Rambow

The paper explores the performance of LLMs in the context of multi-dimensional analytic writing assessments, i.e. their ability to provide both scores and comments based on multiple assessment criteria. Using a corpus of literature reviews written by L2 graduate students and assessed by human experts against 9 analytic criteria, we prompt several popular LLMs to perform the same task under various conditions. To evaluate the quality of feedback comments, we apply a novel feedback comment quality evaluation framework. This framework is interpretable, cost-efficient, scalable, and reproducible, compared to existing methods that rely on manual judgments. We find that LLMs can generate reasonably good and generally reliable multi-dimensional analytic assessments. We release our corpus and code for reproducibility.

COSMMIC: Comment-Sensitive Multimodal Multilingual Indian Corpus for Summarization and Headline Generation

Raghvendra Kumar, Mohammed Salman S A, Aryan Sahu, Tridib Nandi, Pragathi Y P, Sriparna Saha, Jose G Moreno

Despite progress in comment-aware multimodal and multilingual summarization for English and Chinese, research in Indian languages remains limited. This study addresses this gap by introducing COSMMIC, a pioneering comment-sensitive multimodal, multilingual dataset featuring nine major Indian languages. COSMMIC comprises 4,959 article-image pairs and 24,484 reader comments, with ground-truth summaries available in all included languages. Our approach enhances summaries by integrating reader insights and feedback. We explore summarization and headline generation across four configurations: (1) using article text alone, (2) incorporating user comments, (3) utilizing images, and (4) combining text, comments, and images. To assess the dataset's effectiveness, we employ state-of-the-art language models such as LLama3 and GPT-4. We conduct a comprehensive study to evaluate different component combinations, including identifying supportive comments, filtering out noise using a dedicated comment classifier using IndicBERT, and extracting valuable insights from images with a multilingual CLIP-based classifier. This helps determine the most effective configurations for natural language generation (NLG) tasks. Unlike many existing datasets that are either text-only or lack user comments in multimodal settings, COSMMIC uniquely integrates text, images, and user feedback. This holistic approach bridges gaps in Indian language resources, advancing NLP research and fostering inclusivity.

MUSTS: MUltilingual Semantic Textual Similarity Benchmark

Tharindu Ranasinghe, Hansi Hettiarachchi, Constantin Orasan, Ruslan Mitkov

Predicting semantic textual similarity (STS) is a complex and ongoing challenge in natural language processing (NLP). Over the years, researchers have developed a variety of supervised and unsupervised approaches to calculate STS automatically. Additionally, various benchmarks, which include STS datasets, have been established to consistently evaluate and compare these STS methods. However, they largely focus on high-resource languages, mixed with datasets annotated focusing on relatedness instead of similarity and containing automatically translated instances. Therefore, no dedicated benchmark for multilingual STS exists. To solve this gap, we introduce the Multilingual Semantic Textual Similarity Benchmark (MUSTS), which spans 13 languages, including low-resource languages. By evaluating more than 25 models on MUSTS, we establish the most comprehensive benchmark of multilingual STS methods. Our findings confirm that STS remains a challenging task, particularly for low-resource languages.

Does Context Matter? ContextualJudgeBench for Evaluating LLM-based Judges in Contextual Settings

Austin Xu, Srijan Bansal, Yifei Ming, Semih Yavuz, Shafiq Joty

The large language model (LLM)-as-judge paradigm has been used to meet the demand for a cheap, reliable, and fast evaluation of model outputs during AI system development and post-deployment monitoring. While judge models—LLMs finetuned to specialize in assessing and critiquing model outputs—have been touted as general purpose evaluators, they are typically evaluated only on non-contextual scenarios, such as instruction following. The omission of contextual settings—those where external information is used as context to generate an output—is surprising given the increasing prevalence of retrieval-augmented generation (RAG) and summarization use cases. Contextual assessment is uniquely challenging, as evaluation often depends on practitioner priorities, leading to conditional evaluation criteria (e.g., comparing responses based on factuality and then considering completeness if they are equally factual). To address the gap, we propose ContextualJudgeBench, a judge benchmark with 2,000 challenging response pairs across eight splits inspired by real-world contextual evaluation scenarios. We build our benchmark with a multi-pronged data construction pipeline that leverages both existing human annotations and model-based perturbations. Our comprehensive study across 11 judge models and 7 general purpose models, reveals that the contextual information and assessment criteria present a significant challenge to even state-of-the-art models. For example, o1, the best-performing model, barely reaches 55% consistent accuracy.

Polishing Every Facet of the GEM: Testing Linguistic Competence of LLMs and Humans in Korean

SungHo Kim, Nayeon Kim, Taehee Jeon, SangKeun Lee

We introduce the \underline{Ko}rean \underline{G}rammar \underline{E}valuation Bench\underline{M}ark (KoGEM), designed to assess the linguistic competence of LLMs and humans in Korean. KoGEM consists of 1.5k multiple-choice QA pairs covering five main categories and 16 subcategories. The zero-shot evaluation of 27 LLMs of various sizes and types reveals that while LLMs perform remarkably well on straightforward tasks requiring primarily definitional knowledge, they struggle with tasks that demand the integration of real-world experiential knowledge, such as phonological rules and pronunciation. Furthermore, our in-depth analysis suggests that incorporating such experiential knowledge could enhance the linguistic competence of LLMs. With KoGEM, we not only highlight the limitations of current LLMs in linguistic competence but also uncover hidden facets of LLMs in linguistic competence, paving the way for enhancing comprehensive language understanding. Our code and dataset are available at: https://github.com/SungHo3268/KoGEM.

Agri-CM3: A Chinese Massive Multi-modal, Multi-level Benchmark for Agricultural Understanding and Reasoning

Haotian Wang, Yi Guan, Fanshu Meng, Chao Zhao, Lian Yan, Yang Yang, Jingchi Jiang

Multi-modal Large Language Models (MLLMs) integrating images, text, and speech can provide farmers with accurate diagnoses and treatment of pests and diseases, enhancing agricultural efficiency and sustainability. However, existing benchmarks lack comprehensive evaluations, particularly in multi-level reasoning, making it challenging to identify model limitations. To address this issue, we introduce Agri-CM^3, an expert-validated benchmark assessing MLLMs' understanding and reasoning in agricultural management. It includes 3,939 images and 15,901 multi-level multiple-choice questions with detailed explanations. Evaluations of 45 MLLMs reveal significant gaps. Even GPT-4o achieves only 63.64\% accuracy, falling short in fine-grained reasoning tasks. Analysis across three reasoning levels and seven compositional abilities highlights key challenges in accuracy and cognitive understanding. Our study provides insights for advancing MLLMs in agricultural management, driving their development and application. Code and data are available at .

Advancing Zero-shot Text-to-Speech Intelligibility across Diverse Domains via Preference Alignment

Xueyao Zhang, Yuancheng Wang, Chaoren Wang, Ziniu Li, Zhuo Chen, Zhizheng Wu

Modern zero-shot text-to-speech (TTS) systems, despite using extensive pre-training, often struggle in challenging scenarios such as tongue twisters, repeated words, code-switching, and cross-lingual synthesis, leading to intelligibility issues. To address these limitations, this paper leverages preference alignment techniques, which enable targeted construction of out-of-pretraining-distribution data to enhance performance. We introduce a new dataset, named the Intelligibility Preference Speech Dataset (INTP), and extend the Direct Preference Optimization (DPO) framework to accommodate diverse TTS architectures. After INTP alignment, in addition to intelligibility, we observe overall improvements including naturalness, similarity, and audio quality for multiple TTS models across diverse domains. Based on that, we also verify the weak-to-strong generalization ability of INTP for more intelligible models such as CosyVoice 2 and Ints. Moreover, we showcase the potential for further improvements through iterative alignment based on Ints. Audio samples are available at https://intalign.github.io/.

KokoroChat: A Japanese Psychological Counseling Dialogue Dataset Collected via Role-Playing by Trained Counselors

Zhiyang Qi, Takumasa Kaneko, Keiko Takamizo, Mariko Ukiyo, Michimasa Inaba

Generating psychological counseling responses with language models relies heavily on high-quality datasets. Crowdsourced data collection methods require strict worker training, and data from real-world counseling environments may raise privacy and ethical concerns. While recent studies have explored using large language models (LLMs) to augment psychological counseling dialogue datasets, the resulting data often suffers from limited diversity and authenticity. To address these limitations, this study adopts a role-playing approach where trained counselors simulate counselor-client interactions, ensuring high-quality dialogues while mitigating privacy risks. Using this method, we construct KokoroChat, a Japanese psychological counseling dialogue dataset comprising 6,589 long-form dialogues, each accompanied by comprehensive client feedback. Experimental results demonstrate that fine-tuning open-source LLMs with KokoroChat improves both the quality of generated counseling responses and the automatic evaluation of counseling dialogues. The KokoroChat dataset is available at https://github.com/UEC-InabaLab/KokoroChat.

ChildMandarin: A Comprehensive Mandarin Speech Dataset for Young Children Aged 3-5

Jiaming Zhou, shiyao wang, Shiwan Zhao, Jiabei He, Haoqin Sun, Hui Wang, Cheng Liu, Aobo Kong, Yujie Guo, Xi Yang, Yequan Wang, Yonghua Lin, Yong Qin

Automatic speech recognition (ASR) systems have advanced significantly with models like Whisper, Conformer, and self-supervised frameworks such as Wav2vec 2.0 and HuBERT. However, developing robust ASR models for young children's speech remains challenging due to differences in pronunciation, tone, and pace compared to adult speech. In this paper, we introduce a new Mandarin speech dataset focused on children aged 3 to 5, addressing the scarcity of resources in this area. The dataset comprises 41.25 hours of speech with carefully crafted manual transcriptions, collected from 397 speakers across various provinces in China, with balanced gender representation. We provide a comprehensive analysis of speaker demographics, speech duration distribution and geographic coverage. Additionally, we evaluate ASR performance on models trained from scratch, such as Conformer, as well as fine-tuned pre-trained models like HuBERT and Whisper, where fine-tuning demonstrates significant performance improvements. Furthermore, we assess speaker verification (SV) on our dataset, showing that, despite the challenges posed by the unique vocal characteristics of young children, the dataset effectively supports both ASR and SV tasks. This dataset is a valuable contribution to Mandarin child speech research and holds potential for applications in educational technology and child-computer interaction. It will be open-source and freely available for all academic purposes.

Can Knowledge Graphs Make Large Language Models More Trustworthy? An Empirical Study Over Open-ended Question Answering

Yuan Sui, Yufei He, Zifeng Ding, Bryan Hooi

Recent works integrating Knowledge Graphs (KGs) have shown promising improvements in enhancing the reasoning capabilities of Large Language Models (LLMs). However, existing benchmarks primarily focus on closed-ended tasks, leaving a gap in evaluating performance on more complex, real-world scenarios. This limitation also hinders a thorough assessment of KGs' potential to reduce hallucinations in LLMs. To address this, we introduce OKGQA, a new benchmark specifically designed to evaluate LLMs augmented with KGs in open-ended, real-world question answering settings. OKGQA reflects practical complexities through diverse question types and incorporates metrics to quantify both hallucination rates and reasoning improvements in LLM+KG models. To consider the scenarios in which KGs may contain varying levels of errors, we propose a benchmark variant, OKGQA-P, to assess model performance when the semantics and structure of KGs are deliberately perturbed and contaminated. In this paper, we aims to (1) explore whether KGs can make LLMs more trustworthy in an open-ended setting, and (2) conduct a comparative analysis to shed light on method design. We believe this study can facilitate a more complete performance comparison and encourages continuous improvement in integrating KGs with LLMs to mitigate hallucination, and make LLMs more trustworthy.

SDBench: A Survey-based Domain-specific LLM Benchmarking and Optimization Framework

Cheng Guo, Hu Kai, Shuxian Liang, Yiyang Jiang, Yi Gao, Xian-Sheng Hua, Wei Dong

The rapid advancement of large language models (LLMs) in recent years has made it feasible to establish domain-specific LLMs for specialized fields. However, in practical development, acquiring domain-specific knowledge often requires a significant amount of professional expert manpower. Moreover, even when domain-specific data is available, the lack of a unified methodology for benchmark dataset establishment often results in uneven data distribution. This imbalance can lead to an inaccurate assessment of the true model capabilities during the evaluation of domain-specific LLMs. To address these challenges, we introduce **SDBench**, a generic framework for generating evaluation datasets for domain-specific LLMs. This method is also applicable for establishing the LLM instruction datasets. It significantly reduces the reliance on expert manpower while ensuring that the collected data is uniformly distributed. To validate the effectiveness of this framework, we also present the **BridgeBench**, a novel benchmark for bridge engineering knowledge, and the **BridgeGPT**, the first LLM specialized in bridge engineering, which can solve bridge engineering tasks.

Graph-Structured Trajectory Extraction from Travelogues

Aitaro Yamamoto, Hiroyuki Otomo, Hiroki Ouchi, Shohei Higashiyama, Hiroki Teranishi, Hiroyuki Shindo, Taro Watanabe

Human traveling trajectories play a central role in characterizing each travelogue, and automatic trajectory extraction from travelogues is highly desired for tourism services, such as travel planning and recommendation. This work addresses the extraction of human traveling trajectories from travelogues. Previous work treated each trajectory as a sequence of visited locations, although locations with different granularity levels, e.g., "Kyoto City" and "Kyoto Station," should not be lined up in a sequence. In this work, we propose to represent the trajectory as a graph that can capture the hierarchy as well as the visiting order, and construct a benchmark dataset for the trajectory extraction. The experiments using this dataset show that even naive baseline systems can accurately predict visited locations and the visiting order between them, while it is more challenging to predict the hierarchical relations.

MaXIFE: Multilingual and Cross-lingual Instruction Following Evaluation

Yile Liu, Ziwei Ma, Xiu Jiang, Jinglu HU, ChangJing, Liang Li

With the rapid adoption of large language models (LLMs) in natural language processing, the ability to follow instructions has emerged as a key metric for evaluating their practical utility. However, existing evaluation methods often focus on single-language scenarios, overlooking the challenges and differences present in multilingual and cross-lingual contexts. To address this gap, we introduce MaXIFE: a comprehensive evaluation benchmark designed to assess instruction-following capabilities across 23 different languages with 1667 verifiable instruction tasks. MaXIFE integrates both Rule-Based Evaluation and Model-Based Evaluation, ensuring a balance of efficiency and accuracy. We applied MaXIFE to evaluate several leading commercial LLMs, establishing baseline results for future comparisons. By providing a standardized tool for multilingual instruction-following evaluation, MaXIFE aims to advance research and development in natural language processing.

KazMMLU: Evaluating Language Models on Kazakh, Russian, and Regional Knowledge of Kazakhstan

Mukhammed Togmanov, Nurdaulet Mukhituly, Diana Turmakhan, Jonibek Mansurov, Maiya Goloburda, Akhmed Sakip, Zhuohan Xie, Yuxia Wang, Bekassyl Syzdykov, Nurkhan Laiyk, Alham Fikri Aji, Ekaterina Kochmar, Preslav Nakov, Fajri Koto

Despite having a population of twenty million, Kazakhstan's culture and language remain underrepresented in the field of natural language processing. Although large language models (LLMs) continue to advance worldwide, progress in Kazakh language has been limited, as seen in the scarcity of dedicated models and benchmark evaluations. To address this gap, we introduce KazMMLU, the first MMLU-style dataset specifically designed for Kazakh language. KazMMLU comprises 23,000 questions that cover various educational levels, including STEM, humanities, and social sciences, sourced from authentic educational materials and manually validated by native speakers and educators. The dataset includes 10,969 Kazakh questions and 12,031 Russian questions, reflecting Kazakhstan's bilingual education system and rich local context. Our evaluation of several state-of-the-art multilingual models (Llama3.1, Qwen-2.5, GPT-4, and DeepSeek V3) demonstrates substantial room for improvement, as even the best-performing models struggle to achieve competitive performance in Kazakh and Russian. These findings highlight significant performance gaps compared to high-resource languages. We hope that our dataset will enable further research and development of Kazakh-centric LLMs.

Instruction Tuning on Public Government and Cultural Data for Low-Resource Language: a Case Study in Kazakh

Nurkhan Laiyk, Daniil Orel, Rituraj Joshi, Maiya Goloburda, Yuxia Wang, Preslav Nakov, Fajri Koto

Instruction tuning in low-resource languages remains underexplored due to limited text data, particularly in government and cultural domains. To address this, we introduce and open-source a large-scale (10,600 samples) instruction-following (IFT) dataset, covering key institutional and cultural knowledge relevant to Kazakhstan. Our dataset enhances LLMs' understanding of procedural, legal, and structural governance topics. We employ LLM-assisted data generation, comparing open-weight and closed-weight models for dataset construction, and select GPT-4o as the backbone. Each entity of our dataset undergoes full manual verification to ensure high quality. We also show that fine-tuning Qwen, Falcon, and Gemma on our dataset leads to consistent performance improvements in both multiple-choice and generative tasks, demonstrating the potential of LLM-assisted instruction tuning for low-resource languages.

QAEval: Mixture of Evaluators for Question-Answering Task Evaluation

Tan Yue, Rui Mao, xuzhao Shi, SHUO ZHAN, Zuhao Yang, Dongyan Zhao

Question answering (QA) tasks serve as a key benchmark for evaluating generation systems. Traditional rule-based metrics, such as accuracy and relaxed-accuracy, struggle with open-ended and unstructured responses. LLM-based evaluation methods offer greater flexibility but suffer from sensitivity to instructions, robustness issues, and high computational costs. To overcome these challenges, we introduce QAEval, a hybrid framework combining rule-based reliability with LLM-based adaptability. QAEval utilizes two high-quality datasets: QAExtract for short-answer extraction and QAScore for scoring model training. By integrating a Mixture of Evaluators model with Dynamic Load Balancing Optimization, QAEval enables accurate, cost-effective QA evaluation. Experimental results show it outperforms models like GPT-4o and Claude-3, achieving 92.3% accuracy with only 0.6B parameters.

MMMU-Pro: A More Robust Multi-discipline Multimodal Understanding Benchmark

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, Graham Neubig

This paper introduces MMMU-Pro, a robust version of the Massive Multi-discipline Multimodal Understanding and Reasoning (MMMU) benchmark. MMMU-Pro rigorously assesses multimodal models' true understanding and reasoning capabilities through a three-step process based on MMMU: (1) filtering out questions answerable by text-only models, (2) augmenting candidate options, and (3) introducing a vision-only input setting where questions are embedded within images. This setting challenges AI to truly "see" and "read" simultaneously, testing \textit{a core human cognitive skill of seamlessly integrating visual and textual information}. Results show that model performance is substantially lower on MMMU-Pro than on MMMU, ranging from 16.8% to 26.9% across models. We explore the impact of OCR prompts and Chain of Thought (CoT) reasoning, finding that OCR prompts have minimal effect while CoT generally improves performance. MMMU-Pro provides a more rigorous evaluation tool, closely mimicking real-world scenarios and offering valuable directions for future multimodal research.

Cheems: A Practical Guidance for Building and Evaluating Chinese Reward Models from Scratch

Xueru Wen, Jie Lou, Zichao Li, Yaojie Lu, XingYu, Yuqiu Ji, Guohai Xu, Hongyu Lin, Ben He, Xianpei Han, Le Sun, Debing Zhang

Reward models (RMs) are crucial for aligning large language models (LLMs) with human preferences. However, most RM research is centered on English and relies heavily on synthetic resources, which leads to limited and less reliable datasets and benchmarks for Chinese. To address this gap, we introduce CheemsBench, a fully human-annotated RM evaluation benchmark within Chinese contexts, and CheemsPreference, a large-scale and diverse preference dataset annotated through human-machine collaboration to support Chinese RM training. We systematically evaluate open-source discriminative and generative RMs on CheemsBench and observe significant limitations in their ability to capture human preferences in Chinese scenarios. Additionally, based on CheemsPreference, we construct an RM that achieves state-of-the-art performance on CheemsBench, demonstrating the necessity of human supervision in RM training. Our findings reveal that scaled AI-generated data struggles to fully capture human preferences, emphasizing the importance of high-quality human supervision in RM development.

MISP-Meeting: A Real-World Dataset with Multimodal Cues for Long-form Meeting Transcription and Summarization

HangChen, Chao-Han Huck Yang, Jia-Chen Gu, Hongxu Yin, Sabato Marco Siniscalchi, Jun Du

We introduce MISP-Meeting, a new real-world, multimodal dataset that covers subject-oriented long-form content. MISP-Meeting integrates information from speech, vision, and text modalities to facilitate automatic meeting transcription and summarization (AMTS). Challenging conditions in human meetings, including far-field speech recognition, audio-visual understanding, and long-term summarization, have been carefully evaluated. We benchmark state-of-the-art automatic speech recognition (ASR) and large language models (LLMs) on this dataset, enhanced with multimodal cues. Experiments demonstrate that incorporating multimodal cues, such as lip movements and visual focus of attention, significantly enhances transcription accuracy, reducing the character error rate (CER) from 36.60% to 20.27% via guided source separation (GSS), fine-tuning, and audio-visual fusion. Furthermore, our summarization analysis reveals a direct correlation between ASR quality and summary coherence, underscoring the importance of robust multimodal modeling. Our dataset and codebase will be released as open source.

Beyond One-Size-Fits-All: Tailored Benchmarks for Efficient Evaluation

Peiwen Yuan, Yueqi Zhang, Shaoxiong Feng, Yiwei Li, Xinglin Wang, Jiayi Shi, Chuyi Tan, Boyuan Pan, Yao Hu, Kan Li

Evaluating models on large benchmarks can be very resource-intensive, especially during a period of rapid model evolution. Existing efficient evaluation methods estimate the performance of target models by testing them on a small, static coreset derived from the publicly available evaluation results of source models, which are separate from the target models. However, these approaches rely on the assumption that target models have high prediction consistency with source models, which doesn't generalize well in practice. To fill this gap, we propose TailoredBench, a method that conducts customized evaluation tailored to each target model. Specifically, a Global-coreset is first constructed as a probe to identify the most consistent source models for each target model with an adaptive source model selection strategy. Afterwards, a scalable K-Medoids clustering algorithm is proposed to extend the Global-coreset to a tailored Native-coreset for each target model. According to the predictions on respective Native-coreset, we estimate the overall performance of target models with a calibrated estimation strategy. Comprehensive experiments on five benchmarks across over 300 models demonstrate that compared to best performing baselines, TailoredBench achieves an average reduction of 31.4% in MAE of accuracy estimates under the same inference budgets, showcasing strong effectiveness and generalizability.

FinanceReasoning: Make Financial Numerical Reasoning More Credible, Comprehensive, and Challenging

Zichen Tang, Haihong E, Ziyan Ma, Haoyang He, Jiacheng Liu, Zhongjun Yang, Zihua Rong, Rongjin Li, Kun Ji, Huang Qing, Xinyang Hu, Yang Liu, Qianhe Zheng

We introduce **FinanceReasoning**, a novel benchmark designed to evaluate the reasoning capabilities of large reasoning models (LRMs) in financial numerical reasoning problems. Compared to existing benchmarks, our work provides three key advancements. (1) **Credibility**: We update 15.6\% of the questions from four public datasets, annotating 908 new questions with detailed Python solutions and rigorously refining evaluation standards. This enables an accurate assessment of the reasoning improvements of LRMs. (2) **Comprehensiveness**: FinanceReasoning covers 67.8\% of financial concepts and formulas, significantly surpassing existing datasets. Additionally, we construct 3,133 Python-formatted functions, which enhances LRMs' financial reasoning capabilities through refined knowledge (*e.g.*, 83.2\% \rightarrow 91.6\% for GPT-4o). (3) **Challenge**: Models are required to apply multiple financial formulas for precise numerical reasoning on 238 *Hard* problems. The best-performing model (*i.e.*, OpenAI o1 with PoT) achieves 89.1\% accuracy, yet LRMs still face challenges in numerical precision. We demonstrate that combining Reasoner and Programmer models can effectively enhance LRMs' performance (*e.g.*, 83.2\% \rightarrow 87.8\% for DeepSeek-R1). Our work paves the way for future research on evaluating and improving LRMs in domain-specific complex reasoning tasks.

Program Synthesis Benchmark for Visual Programming in XLogoOnline Environment

Chao Wen, Jacqueline Staub, Adish Singla

Large language and multimodal models have shown remarkable success on various benchmarks focused on specific skills such as general-purpose programming, math word problem-solving, and visual question answering. However, it is unclear how well these models perform on tasks that require a combination of these skills. In this paper, we curate a novel program synthesis benchmark based on the real-world tasks in the XLogoOnline visual programming environment. Each task requires a combination of different skills such as spatial planning, basic programming, and logical reasoning. Our evaluation shows that current state-of-the-art models like GPT-4V and Llama3-70B struggle to solve these tasks, achieving only 20% and 2.35% success rates, respectively. Next, we develop a fine-tuning pipeline to boost the performance of models by leveraging a large-scale synthetic training dataset with over 80,000 tasks. Moreover, we showcase how emulator-driven feedback can be used to design a curriculum over training data distribution, through which a fine-tuned Llama3-8B drastically outperforms GPT-4V and Llama3-70B models. Finally, we provide an in-depth failure analysis to understand the limitations of different models. We will publicly release the benchmark for future research on program synthesis in visual programming.

Can external validation tools improve annotation quality for LLM-as-a-Judge?

Arduin Findeis, Floris Weers, Guoli Yin, Ke Ye, Ruoming Pang, Tom Gunter

Pairwise preferences over model responses are widely collected to evaluate and provide feedback to large language models (LLMs). Given two alternative model responses to the same input, a human or AI annotator selects the “better” response. This approach can provide feedback for domains where other hard-coded metrics are difficult to obtain (e.g., chat response quality), thereby helping model evaluation or training. However, for some domains high-quality pairwise comparisons can be tricky to obtain - from AI and humans. For example, for responses with many factual statements, annotators may disproportionately weigh writing quality rather than underlying facts. In this work, we explore augmenting standard AI annotator systems with additional tools to improve performance on three challenging response domains: long-form factual, math and code tasks. We propose a tool-using agentic system to provide higher quality feedback on these domains. Our system uses web-search and code execution to ground itself based on external validation, independent of the LLM’s internal knowledge and biases. We provide extensive experimental results evaluating our method across the three targeted response domains as well as general annotation tasks, using RewardBench (incl. AlpacaEval and LLMBar), RewardMath, as well as three new datasets for domains with saturated pre-existing datasets. Our results indicate that external tools can indeed improve performance in many, but not all, cases. More generally, our experiments highlight the sensitivity of performance to simple parameters (e.g., prompt) and the need for improved (non-saturated) annotator benchmarks. We share our code at https://github.com/apple/ml-agent-evaluator.

Com^2 : A Causal-Guided Benchmark for Exploring Complex Commonsense Reasoning in Large Language Models

Kai Xiong, Xiao Ding, Yixin Cao, Yuxiong Yan, Li Du, Yufei Zhang, Jinglong Gao, Jiaqian Liu, Bing Qin, Ting Liu

Large language models (LLMs) have mastered abundant simple and explicit commonsense knowledge through pre-training, enabling them to achieve human-like performance in simple commonsense reasoning. Nevertheless, LLMs struggle to reason with complex and implicit commonsense knowledge that is derived from simple ones (such as understanding the long-term effects of certain events), an aspect humans tend to focus on more. Existing works focus on complex tasks like math and code, while complex commonsense reasoning remains underexplored due to its uncertainty and lack of structure. To fill this gap and align with real-world concerns, we propose a benchmark Com^2 focusing on complex commonsense reasoning. We first incorporate causal event graphs to serve as structured complex commonsense. Then we adopt causal theory (e.g., intervention) to modify the causal event graphs and obtain different scenarios that meet human concerns. Finally, an LLM is employed to synthesize examples with slow thinking, which is guided by the logical relationships in the modified causal graphs. Furthermore, we use detective stories to construct a more challenging subset. Experiments show that LLMs struggle in reasoning depth and breadth, while post-training and slow thinking can alleviate this. The code and data are available at https://github.com/Waste-Wood/Com2}.

CritiQ: Mining Data Quality Criteria from Human Preferences

Honglin Guo, Kai Lv, Qipeng Guo, Tianyi Liang, Zhiheng Xi, Demin Song, Qiuyinzhe Zhang, Yu Sun, Kai Chen, Xipeng Qiu, Tao Gui

Language model heavily depends on high-quality data for optimal performance. Existing approaches rely on manually designed heuristics, the perplexity of existing models, training classifiers, or careful prompt engineering, which require significant expert experience and human annotation effort while introduce biases. We introduce CritiQ, a novel data selection method that automatically mines criteria from human preferences for data quality with only \sim30 human-annotated pairs and performs efficient data selection. The main component, CritiQ Flow, employs a manager agent to evolve quality criteria and worker agents to make pairwise judgments. We build a knowledge base that extracts quality criteria from previous work to boost CritiQ Flow. Compared to perplexity- and classifier-based methods, verbal criteria are more interpretable and have greater reusable value. After deriving the criteria, we train the CritiQ Scorer to give quality scores and perform efficient data selection. We demonstrate the effectiveness of our method in the code, math, and logic domains, achieving high accuracy on human-annotated test sets. To validate the quality of the selected data, we continually train Llama 3.2 models and observe improved performance on downstream tasks compared to uniform sampling. Ablation studies validate the benefits of the knowledge base and the reflection process. We analyze how criteria evolve and the effectiveness of majority voting.

MockConf: A Student Interpretation Dataset: Analysis, Word- and Span-level Alignment and Baselines

Dávid Javorský, Ondřej Bojar, François Yvon

In simultaneous interpreting, an interpreter renders the speech into another language with a very short lag, much sooner than sentences are finished. In order to understand and later reproduce this dynamic and complex task automatically, we need specialized datasets and tools for analysis, monitoring, and evaluation, such as parallel speech corpora, and tools for their automatic annotation. Existing parallel corpora of translated texts and associated alignment algorithms hardly fill this gap, as they fail to model long-range interactions between speech segments or specific types of divergences (e.g. shortening, simplification, functional generalization) between the original and interpreted speeches. In this work, we develop and explore MockConf, a student interpretation dataset that was collected from Mock Conferences run as part of the students' curriculum. This dataset contains 7hours of recordings in 5 European languages, transcribed and aligned at the level of spans and words. We further implement and release InterAlign, a modern web-based annotation tool for parallel word and span annotations on long inputs, suitable for aligning simultaneous interpreting. We propose metrics for the evaluation and a baseline for automatic alignment. Dataset and tools will be released to the community.

CalibraEval: Calibrating Prediction Distribution to Mitigate Selection Bias in LLMs-as-Judges

Haitao Li, Junjie Chen, Qingyao Ai, Zhumin Chu, Yujia Zhou, Qian Dong, Yiqun LIU

The use of large language models (LLMs) as automated evaluation tools to assess the quality of generated natural language, known as "LLMs-as-Judges", has demonstrated promising capabilities and is rapidly gaining widespread attention. However, when applied to pairwise comparisons of candidate responses, LLM-based evaluators often exhibit selection bias. Specifically, their judgments may become inconsistent when the option positions or ID tokens are swapped, compromising the effectiveness and fairness of the evaluation result. To address this challenge, we introduce CalibraEval, a novel label-free method for mitigating selection bias during inference. Specifically, CalibraEval reformulates debiasing as an optimization task aimed at adjusting observed prediction distributions to align with unbiased prediction distributions. To solve this optimization problem, we propose a non-parametric order-preserving algorithm (NOA). This algorithm leverages the partial order relationships between model prediction distributions, thereby eliminating the need for explicit labels and precise mathematical function modeling. Empirical evaluations of LLMs in multiple representative benchmarks demonstrate that CalibraEval effectively mitigates selection bias and improves performance compared to existing debiasing methods. This work marks a step toward building more robust and unbiased automated evaluation frameworks, paving the way for improved reliability in AI-driven assessments. The code can be found at https://github.com/CSHaitao/CalibraEval.

Global Eye: Breaking the "Fixed Thinking Pattern" during the Instruction Expansion Process

wenxuan lu, Wei Liu, Jian Luan, Bin Wang, Songhao Jiang, Tianning Zang

An extensive high-quality instruction dataset is crucial for the instruction tuning process of Large Language Models (LLMs). Recent instruction expansion methods have demonstrated their capability to improve the quality and quantity of existing datasets, by prompting high-performance LLM to generate multiple new instructions from the original ones. However, existing methods focus on constructing multi-perspective prompts (e.g., increasing complexity or difficulty) to expand instructions, overlooking the "Fixed Thinking Pattern" issue of LLMs. This issue arises when repeatedly using the same set of prompts, causing LLMs to rely on a limited set of certain expressions to expand all instructions, potentially compromising the diversity of the final expanded dataset. This paper theoretically analyzes the causes of the "Fixed Thinking Pattern", and corroborates this phenomenon through multi-faceted empirical research. Furthermore, we propose a novel method based on dynamic prompt updating: Global Eye. Specifically, after a fixed number of instruction expansions, we analyze the statistical characteristics of newly generated instructions and then update the prompts. Experimental results show that our method enables Llama3-8B and Llama2-13B to surpass the performance of open-source LLMs and GPT3.5 across various metrics. Our code and data are submitted to the Software \& Data option.

Mis-prompt: Benchmarking Large Language Models for Proactive Error Handling

Jiayi Zeng, Yizhe Feng, Mengliang He, Wenhui Lei, Wei Zhang, Zeming Liu, Xiaoming Shi, Aimin Zhou

Large language models (LLMs) have demonstrated significant advancements in error handling. Current error-handling works are performed in a passive manner, with explicit error-handling instructions. However, in real-world scenarios, explicit error-handling instructions are usually unavailable. In this paper, our work identifies this challenge as how to conduct proactive error handling without explicit error handling instructions. To promote further research, this work introduces a new benchmark, termed Mis-prompt, consisting of four evaluation tasks, an error category taxonomy, and a new evaluation dataset. Furthermore, this work analyzes current LLMs' performance on the benchmark, and the experimental results reveal that current LLMs show poor performance on proactive error handling, and SFT on error handling instances improves LLMs' proactive error handling capabilities. The dataset will be publicly available.

TripCraft: A Benchmark for Spatio-Temporally Fine Grained Travel Planning

Soumyabrata Chaudhuri, Pranav Purkar, Ritwik Raghav, Shubhojit Mallick, Manish Gupta, Abhik Jana, Shreya Ghosh

Recent advancements in probing Large Language Models (LLMs) have explored their latent potential as personalized travel planning agents, though this remains a rather nascent field. Existing benchmarks, such as TravelPlanner and TravelPlanner+, rely on semi-synthetic data as well ignoring several key components of travel planning, limiting their real-world applicability. Therefore, we introduce TripCraft, a spatio-temporally coherent travel planning dataset incorporating real-world constraints, including public transit schedules, public events, varied attraction categories, and user personas for enhanced personalization. Our dataset enables more detailed trip itinerary generation (including duration spent at each point of interest based on users' persona, transit between two points of interest, etc.) while ensuring spatio-temporal consistency. Further, we propose novel evaluation metrics (temporal meal score, attraction score, spatial score, ordering score, and persona score) to assess LLM-generated plans across temporal, spatial, sequential, and personal dimensions, overcoming the limitations of commonsense and hard constraint metrics. Interestingly, our parameter-informed setting significantly enhances meal scheduling, improving performance from 61% to 80% in the 7-day scenario- as quantified by a 19% gain in our temporal meal score. Moreover, TripCraft serves as a high-quality benchmark for advancing personalized LLM-driven travel planning.

SCOP: Evaluating the Comprehension Process of Large Language Models from a Cognitive View

Yongjie Xiao, Hongru Liang, Peixin Qin, YAO ZHANG, Wenqiang Lei

Despite the great potential of large language models(LLMs) in machine comprehension, it is still disturbing to fully count on them in real-world scenarios. This is probably because there is no rational explanation for whether the comprehension process of LLMs is aligned with that of experts. In this paper, we propose SCOP to carefully examine how LLMs perform during the comprehension process from a cognitive view. Specifically, it is equipped with a systematical definition of five requisite skills during the comprehension process, a strict framework to construct testing data for these skills, and a detailed analysis of advanced open-sourced and closed-sourced LLMs using the testing data. With SCOP, we find that it is still challenging for LLMs to perform an expert-level comprehension process. Even so, we notice that LLMs share some similarities with experts, e.g., performing better at comprehending local information than global information. Further analysis reveals that LLMs can be somewhat unreliable --- they might reach correct answers through flawed comprehension processes. Based on SCOP, we suggest that one direction for improving LLMs is to focus more on the comprehension process, ensuring all comprehension skills are thoroughly developed during training.

Revisit Self-Debugging with Self-Generated Tests for Code Generation

Xiancai Chen, Zhengwei Tao, Kechi Zhang, Changzhi Zhou, Xinyu Zhang, Wanli Gu, Yuanpeng He, Mengdi Zhang, Xunliang Cai, Haiyan Zhao, Zhi Jin

Large language models (LLMs) have demonstrated significant advancements in code generation, yet they still face challenges when tackling tasks that extend beyond their basic capabilities. Recently, the concept of self-debugging has been proposed as a way to enhance code generation performance by leveraging execution feedback from tests. However, the availability of high-quality tests in real-world scenarios is often limited. In this context, self-debugging with self-generated tests emerges as a promising solution, though its limitations and practical potential have not been fully explored. To address this gap, we investigate the efficacy of self-debugging in code generation tasks. We propose and analyze two distinct paradigms for the self-debugging process: post-execution and in-execution self-debugging. Our findings reveal that post-execution self-debugging struggles with the test bias introduced by self-generated tests, which can lead to misleading feedback. In contrast, in-execution self-debugging enables LLMs to mitigate this bias and leverage intermediate states during program execution. By focusing on runtime information rather than relying solely on potentially flawed self-generated tests, this approach demonstrates significant promise for improving the robustness and accuracy of LLMs in code generation tasks.

Crowdsource, Crawl, or Generate? Creating SEA-VL, a Multicultural Vision-Language Dataset for Southeast Asia

Samuel Cahyawijaya, Holy Lovenia, Joel Ruben Antony Moniz, Tack Hwa Wong, Mohammad Rifqi Farhansyah, Thant Thiri Maung, Frederikus Hudi, David Anugraha, Muhammad Ravi Shulthan Habibi, Muhammad Reza Qorib, Amit Agarwal, Joseph Marvin Imperial, Hitesh Laxmichand Patel, Vicky Feliren, Bahrul Ilmi Nasution, Manuel Antonio Rufino, Genta Indra Winata, Rian Adam Rajagede, Carlos Rafael Catalan, Mohamed Fazli Mohamed Imam, Priyaranjan Pattnayak, Salsabila Zahirah Pranida, Kevin Pratama, Yeshil Bangera, Adisai Na-Thalang, Patricia Nicole Monderin, Yueqi Song, christian simon, Lynnette Hui Xian Ng, Richardy Lobo Sapan, Taki Hasan Rafi, Bin Wang, Supryadi, Kanyakorn Veerakanjana, Piyalitt Ittichaiwong, Matthew Theodore Roque, Karissa Vincentio, Takdanai Kreangphet, Phakphum Artkaew, Kadek Hendrawan Palgunadi, Yanzhi Yu, Rochana Prih Hastuti, William Nixon, Mithil Bangera, Adrian Xuan Wei Lim, Aye Hninn Khine, Hanif Muhammad Zhafran, Teddy Ferdinan, Audra Aurora Izzani, Ayushman Singh, Evan, Jauza Akbar Krito, Michael Anugraha, Fenal Ashokbhai Ilasariya, Haochen Li, John Amadeo Daniswara, Filbert Aurelian Tjiaranata, Eryawan Presma Yulianrifat,

Can Udomcharoenchaikit, Fadil Risdian Ansori, Mahardika Krisna Ihsani, Giang Nguyen, Anab Maulana Barik, Dan John Velasco, Rifo Ahmad Genadi, Saptarshi Saha, Chengwei Wei, Isaiah Edri W. Flores, Kenneth Chen Ko Han, Anjela Gail D. Santos, Wan Shen Lim, Kaung Si Phyo, Tim Santos, Meisyarah Dwiastuti, Jiayun Luo, Jan Christian Blaise Cruz, Ming Shan Hee, Ikhlasul Akmal Hanif, M.Alif Al Hakim, Muhammad Rizky Sya'ban, Kun Kerdthaisong, Lester James Validad Miranda, Fajri Koto, Tirana Noor Fatyanosa, Alham Fikri Aji, Jostin Jerico Rosal, Jun Kevin, Robert Wijaya, Onno P. Kampman, Ruochen Zhang, Börje F. Karlsson, Peerat Limkonchotiwat

Despite Southeast Asia's (SEA) extraordinary linguistic and cultural diversity, the region remains significantly underrepresented in vision-language (VL) research, resulting in AI models that inadequately capture SEA cultural nuances. To fill this gap, we present SEA-VL, an open-source initiative dedicated to developing culturally relevant high-quality datasets for SEA languages. By involving contributors from SEA countries, SEA-VL ensures better cultural relevance and diversity, fostering greater inclusivity of underrepresented languages and cultural depictions in VL research. Our methodology employed three approaches: community-driven crowdsourcing with SEA contributors, automated image crawling, and synthetic image generation. We evaluated each method's effectiveness in capturing cultural relevance. We found that image crawling achieves approximately 85% cultural relevance while being more cost- and time-efficient than crowdsourcing, whereas synthetic image generation failed to accurately reflect SEA cultural nuances and contexts. Collectively, we gathered 1.28 million SEA culturally relevant images, more than 50 times larger than other existing datasets. This work bridges the representation gap in SEA, establishes a foundation for developing culturally aware AI systems for this region, and provides a replicable framework for addressing representation gaps in other underrepresented regions.

Words of Warmth: Trust and Sociability Norms for over 26k English Words

Saif M. Mohammad

Social psychologists have shown that Warmth (W) and Competence (C) are the primary dimensions along which we assess other people and groups. These dimensions impact various aspects of our lives from social competence and emotion regulation to success in the work place and how we view the world. More recent work has started to explore how these dimensions develop, why they have developed, and what they constitute. Of particular note, is the finding that warmth has two distinct components: Trust (T) and Sociability (S). In this work, we introduce Words of Warmth, the first large-scale repository of manually derived word--warmth (as well as word--trust and word--sociability) associations for over 26k English words. We show that the associations are highly reliable. We use the lexicons to study the rate at which children acquire WCTS words with age. Finally, we show that the lexicon enables a wide variety of bias and stereotype research through case studies on various target entities. Words of Warmth is freely available at: http://saifmohammad.com/warmth.html

A Reality Check on Context Utilisation for Retrieval-Augmented Generation

Lovisa Hagström, Sara Vera Marjanovic, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, Isabelle Augenstein

Retrieval-augmented generation (RAG) helps address the limitations of parametric knowledge embedded within a language model (LM). In real world settings, retrieved information can vary in complexity, yet most investigations of LM utilisation of context has been limited to synthetic text. We introduce DRUID (Dataset of Retrieved Unreliable, Insufficient and Difficult-to-understand contexts) with real-world queries and contexts manually annotated for stance. The dataset is based on the prototypical task of automated claim verification, for which automated retrieval of real-world evidence is crucial. We compare DRUID to synthetic datasets (CounterFact, ConflictQA) and find that artificial datasets often fail to represent the complexity and diversity of realistically retrieved context. We show that synthetic datasets exaggerate context characteristics rare in real retrieved data, which leads to inflated context utilisation results, as measured by our novel ACU score. Moreover, while previous work has mainly focused on singleton context characteristics to explain context utilisation, correlations between singleton context properties and ACU on DRUID are surprisingly small compared to other properties related to context source. Overall, our work underscores the need for real-world aligned context utilisation studies to represent and improve performance in real-world RAG settings.

WE-MATH: Does Your Large Multimodal Model Achieve Human-like Mathematical Reasoning?

Runqi Qiao, Qiuna Tan, Guanting Dong, MinhuiWu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma GongQue, Shanglin Lei, YiFan Zhang, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Xiao Zong, Yida Xu, Peiqing Yang, Zhimin Bao, Muxi Diao, Chen Li, Honggang Zhang

Visual mathematical reasoning, as a fundamental visual reasoning ability, has received widespread attention from the Large Multimodal Models (LMMs) community. Existing benchmarks mainly focus more on the end-to-end performance, but neglect the underlying principles of knowledge acquisition and generalization. Instead, we introduce WE-MATH, the first benchmark specifically designed to explore the problem-solving principles. We meticulously collect 6.5K visual math problems and decompose them into 10.9K step-level questions for evaluation, spanning 5 layers of knowledge granularity and 67 hierarchical knowledge concepts. Specifically, we decompose composite problems into sub-problems according to the required knowledge concepts and introduce a novel four-dimensional metric to hierarchically assess inherent issues in LMMs' reasoning process. With WE-MATH, we conduct a thorough evaluation of existing LMMs in visual mathematical reasoning and provide comprehensive analysis and insight for future development. We anticipate that WE-MATH will open new pathways for advancements in visual mathematical reasoning for LMMs. Data and code are available at https://github.com/We-Math/We-Math.

CORDIAL: Can Multimodal Large Language Models Effectively Understand Coherence Relationships?

Aashish Anantha Ramakrishnan, Aadarsh Anantha Ramakrishnan, Dongwon Lee

Multimodal Large Language Models (MLLMs) are renowned for their superior instruction-following and reasoning capabilities across diverse problem domains. However, existing benchmarks primarily focus on assessing factual and logical correctness in downstream tasks, with limited emphasis on evaluating MLLMs' ability to interpret pragmatic cues and intermodal relationships. To address this gap, we assess the competency of MLLMs in performing Multimodal Discourse Analysis (MDA) using Coherence Relations. Our benchmark, CORDIAL, encompasses a broad spectrum of Coherence Relations across 3 different discourse domains at varying levels of granularity. Through our experiments on 10+ MLLMs employing different prompting strategies, we show that even top models like Gemini 1.5 Pro and GPT-4o fail to match the performance of simple classifier-based baselines. This study emphasizes the need to move beyond similarity-based metrics and adopt a discourse-driven framework for evaluating MLLMs, providing a more nuanced assessment of their capabilities. The benchmark and code are available at: https://aashish2000.github.io/CORDIAL/.

LLM Meets Scene Graph: Can Large Language Models Understand and Generate Scene Graphs? A Benchmark and Empirical Study

Dongil Yang, Minjin Kim, Sunghwan Kim, Beong-woo Kwak, Minjun Park, Jinseok Hong, Woontack Woo, Jinyoung Yeo

The remarkable reasoning and generalization capabilities of Large Language Models (LLMs) have paved the way for their expanding applications in embodied AI, robotics, and other real-world tasks. To effectively support these applications, grounding in spatial and temporal understanding in multimodal environments is essential. To this end, recent works have leveraged scene graphs, a structured representation that encodes entities, attributes, and their relationships in a scene. However, a comprehensive evaluation of LLMs' ability to utilize scene graphs remains limited. In this work, we introduce Text-Scene Graph (TSG) Bench, a benchmark designed to systematically assess LLMs' ability to (1) understand scene graphs and (2) generate them from textual narratives. With TSG Bench we evaluate 11 LLMs and reveal that, while models perform well on scene graph understanding, they struggle with scene graph generation, particularly for complex narratives. Our analysis indicates that these models fail to effectively decompose discrete scenes from a complex narrative, leading to a bottleneck when generating scene graphs. These findings underscore the need for improved methodologies in scene graph generation and provide valuable insights for future research. The demonstration of our benchmark is available at https://tsg-bench.netlify.app. Additionally, our code and evaluation data are publicly available at https://github.com/docworlds/tsg-bench.

IOPO: Empowering LLMs with Complex Instruction Following via Input-Output Preference Optimization

Xinghua Zhang, Haiyang Yu, ChengFu, Fei Huang, Yongbin Li

In the realm of large language models (LLMs), the ability of models to accurately follow instructions is paramount as more agents and applications leverage LLMs for construction, where the complexity of instructions are rapidly increasing. However, on the one hand, there is only a certain amount of complex instruction evaluation data; on the other hand, there are no dedicated algorithms to improve the ability to follow complex instructions. To this end, this paper introduces Trace, a benchmark for improving and evaluating the complex instruction-following ability, which consists of 120K training data and 1K evaluation data. Furthermore, we propose IOPO (Input-Output Preference Optimization) alignment method which takes both input and output preference pairs into consideration, where LLMs not only rapidly align with response preferences but also meticulously explore the instruction preferences. Extensive experiments on both in-domain and out-of-domain datasets confirm the effectiveness of IOPO, showing 8.15%, 2.18% improvements on in-domain data and 5.91%, 2.83% on out-of-domain data compared to SFT and DPO respectively. Our code and dataset are released at https://anonymous.4open.science/r/Code7-34A5.

The Essence of Contextual Understanding in Theory of Mind: A Study on Question Answering with Story Characters

Chulun Zhou, Qiujing Wang, Mo Yu, Xiaoqian Yue, Rui Lu, Jiangnan Li, Yifan Zhou, Shunchi Zhang, Jie Zhou, Wai Lam

Theory-of-Mind (ToM) is a fundamental psychological capability that allows humans to understand and interpret the mental states of others. Humans infer others' thoughts by integrating causal cues and indirect clues from broad contextual information, often derived from past interactions. In other words, human ToM heavily relies on the understanding about the backgrounds and life stories of others. Unfortunately, this aspect is largely overlooked in existing benchmarks for evaluating machines' ToM capabilities, due to their usage of short narratives without global context, especially personal background of characters. In this paper, we verify the importance of comprehensive contextual understanding about personal backgrounds in ToM and assess the performance of LLMs in such complex scenarios. To achieve this, we introduce CharToM-QA benchmark, comprising 1,035 ToM questions based on characters from classic novels. Our human study reveals a significant disparity in performance: the same group of educated participants performs dramatically better when they have read the novels compared to when they have not. In parallel, our experiments on state-of-the-art LLMs, including the very recent o1 and DeepSeek-R1 models, show that LLMs still perform notably worse than humans, despite that they have seen these stories during pre-training. This highlights the limitations of current LLMs in capturing the nuanced contextual information required for ToM reasoning.

Beyond the Answer: Advancing Multi-Hop QA with Fine-Grained Graph Reasoning and Evaluation

Qichuan Liu, Chentao Zhang, Chenfeng Zheng, Guosheng Hu, Xiaodong Li, Zhihong Zhang

Recent advancements in large language models (LLMs) have significantly improved the performance of multi-hop question answering (MHQA) systems. Despite the success of MHQA systems, the evaluation of MHQA is not deeply investigated. Existing evaluations mainly focus on comparing the final answers of the reasoning method and given ground-truths. We argue that the reasoning process should also be evaluated because wrong reasoning process can also lead to the correct final answers. Motivated by this, we propose a "Planner-Executor-Reasoner" (PER) architecture, which forms the core of the Plan-anchored Data Preprocessing (PER-DP) and the Plan-guided Multi-Hop QA (PER-QA). The former provides the ground-truth of intermediate reasoning steps and final answers, and the latter offers them of a reasoning method. Moreover, we design a fine-grained evaluation metric called Plan-aligned Stepwise Evaluation (PSE), which evaluates the intermediate reasoning steps from two aspects: planning and solving. Extensive experiments on ten types of questions demonstrate competitive reasoning performance, improved explainability of the MHQA system, and uncover issues such as "fortuitous reasoning continuance" and "latent reasoning suspension" in RAG-based MHQA systems. Besides, we also demonstrate the potential of our approach in data contamination scenarios.

QualiSpeech: A Speech Quality Assessment Dataset with Natural Language Reasoning and Descriptions

Siyin Wang, Wenyi Yu, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Lu Lu, Yu Tsao, Junichi Yamagishi, Yuxuan Wang, Chao Zhang

This paper explores a novel perspective to speech quality assessment by leveraging natural language descriptions, offering richer, more nuanced insights than traditional numerical scoring methods. Natural language feedback provides instructive recommendations and detailed evaluations, yet existing datasets lack the comprehensive annotations needed for this approach. To bridge this gap, we introduce QualiSpeech, a comprehensive low-level speech quality assessment dataset encompassing 11 key aspects and detailed natural language comments that include reasoning and contextual insights. Additionally, we propose the QualiSpeech Benchmark to evaluate the low-level speech understanding capabilities of auditory large language models (LLMs). Experimental results demonstrate that finetuned auditory LLMs can reliably generate detailed descriptions of noise and distortion, effectively identifying their types and temporal characteristics. The results further highlight the potential for incorporating reasoning to enhance the accuracy and reliability of quality assessments. The dataset can be found at https://huggingface.co/datasets/tsinghua-ee/QualiSpeech.

Sandcastles in the Storm: Revisiting the (Im)possibility of Strong Watermarking

Fabrice Y Harel-Canada, Boran Erol, Connor Choi, Jason Liu, Gary Jiarui Song, Nanyun Peng, Amit Sahai

Watermarking AI-generated text is critical for combating misuse. Yet recent theoretical work argues that any watermark can be erased via random walk attacks that perturb text while preserving quality. However, such attacks rely on two key assumptions: (1) rapid mixing (watermarks dissolve quickly under perturbations) and (2) reliable quality preservation (automated quality oracles perfectly guide edits). Through large-scale experiments and human-validated assessments, we find mixing is slow: 100% of perturbed texts retain traces of their origin after hundreds of edits, defying rapid mixing. Oracles falter, as state-of-the-art quality detectors misjudge edits (77% accuracy), compounding errors during attacks. Ultimately, attacks underperform: automated walks remove watermarks just 26% of the time -- dropping to 10% under human quality review. These findings challenge the inevitability of watermark removal. Instead, practical barriers -- slow mixing and imperfect quality control -- reveal watermarking to be far more robust than theoretical models suggest. The gap between idealized attacks and real-world feasibility underscores the need for stronger watermarking methods and more realistic attack models.

WinSpot: GUI Grounding Benchmark with Multimodal Large Language Models

Zheng Hui, Yinheng Li, Dan Zhao, Colby Banbury, Tianyi Chen, Kazuhito Koishida

Graphical User Interface (GUI) automation relies on accurate GUI grounding. However, obtaining large-scale, high-quality labeled data remains a key challenge, particularly in desktop environments like Windows Operating System (OS). Existing datasets primarily focus on structured web-based elements, leaving a gap in real-world GUI interaction data for non-web applications. To address this, we introduce a new framework that leverages LLMs to generate large-scale GUI grounding data, enabling automated and scalable labeling across diverse interfaces. To ensure high accuracy and reliability, we manually validated and refined 5,000 GUI coordinate-instruction pairs, creating WinSpot—the first benchmark specifically designed for GUI grounding tasks in Windows environments. WinSpot provides a high-quality dataset for training and evaluating visual GUI agents, establishing a foundation for future research in GUI automation across diverse and unstructured desktop environments.

HintsOfTruth: A Multimodal Checkworthiness Detection Dataset with Real and Synthetic Claims

Michiel van der Meer, Pavel Korshunov, Sébastien Marcel, Lonneke van der Plas

Misinformation can be countered with fact-checking, but the process is costly and slow. Identifying checkworthy claims is the first step, where automation can help scale fact-checkers' efforts. However, detection methods struggle with content that is (1) multimodal, (2) from diverse domains, and (3) synthetic. We introduce HintsOfTruth, a public dataset for multimodal checkworthiness detection with 27K real-world and synthetic image/claim pairs. The mix of real and synthetic data makes this dataset unique and ideal for benchmarking detection methods. We compare fine-tuned and prompted Large Language Models (LLMs). We find that well-configured lightweight text-based encoders perform comparably to multimodal models but the former only focus on identifying non-claim-like content. Multimodal LLMs can be more accurate but come at a significant computational cost, making them impractical for large-scale applications. When faced with synthetic data, multimodal models perform more robustly.

BeaverTails v2: Towards Multi-Level Safety Alignment for LLMs with Human Preference

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Qiu, Jiayi Zhou, Kaile Wang, Boxun Li, Sirui Han, Yike Guo, Yaodong Yang

In this work, we introduce the PKU-SafeRLHF dataset, designed to promote research on safety alignment in large language models (LLMs). As a sibling project to SafeRLHF and BeaverTails, we separate annotations of helpfulness and harmlessness for question-answering pairs, providing distinct perspectives on these coupled attributes. Overall, we provide 44.6k refined prompts and 265k question-answer pairs with safety meta-labels for 19 harm categories and three severity levels ranging from minor to severe, with answers generated by Llama-family models. Based on this, we collected 166.8k preference data, including dual-preference (helpfulness and harmlessness decoupled) and single-preference data (trade-off the helpfulness and harmlessness from scratch), respectively. Using the large-scale annotation data, we further train severity-sensitive moderation for the risk control of LLMs and safety-centric RLHF algorithms for the safety alignment of LLMs. We believe this dataset will be a valuable resource for the community, aiding in the safe deployment of LLMs.

Robust Estimation of Population-Level Effects in Repeated-Measures NLP Experimental Designs

Alejandro Benito-Santos, Adrian Ghajari, Víctor Fresno

NLP research frequently grapples with multiple sources of variability—spanning runs, datasets, annotators, and more—yet conventional analysis methods often neglect these hierarchical structures, threatening the reproducibility of findings. To address this gap, we contribute a case study illustrating how linear mixed-effects models (LMMs) can rigorously capture systematic language-dependent differences (i.e., population-level effects) in a population of monolingual and multilingual language models. In the context of a bilingual hate speech detection task, we demonstrate that LMMs can uncover significant population-level effects—even under low-resource (small-N) experimental designs—while mitigating confounds and random noise. By setting out a transparent blueprint for repeated-measures experimentation, we encourage the NLP community to embrace variability as a feature, rather than a nuisance, in order to advance more robust, reproducible, and ultimately trustworthy results.

Session 7: Resources and Evaluation (Tuesday 29 July, 10:30-12:00, Hall 5X)

MapQaTor: An Extensible Framework for Efficient Annotation of Map-Based QA Datasets

Mahir Labib Dihan, Mohammed Eunus Ali, Md Rizwan Parvez

Mapping and navigation services like Google Maps, Apple Maps, OpenStreetMap, are essential for accessing various location-based data, yet they often struggle to handle natural language geospatial queries. Recent advancements in Large Language Models (LLMs) show promise in question answering (QA), but creating reliable geospatial QA datasets from map services remains challenging. We introduce MapQaTor, an extensible open-source framework that streamlines the creation of reproducible, traceable map-based QA datasets. MapQaTor enables seamless integration with any maps API, allowing users to gather and visualize data from diverse sources with minimal setup. By caching API responses, the platform ensures consistent ground truth, enhancing the reliability of the data even as real-world information evolves. MapQaTor centralizes data retrieval, annotation, and visualization within a single platform, offering a unique opportunity to evaluate the current state of LLM-based geospatial reasoning while advancing their capabilities for improved geospatial understanding. Evaluation metrics show that, MapQaTor speeds up the annotation process by at least 30 times compared to manual methods, underscoring its potential for developing geospatial resources, such as complex map reasoning datasets. The website is live at: https://mapqator.github.io/ and a demo video is available at: https://youtu.be/bVv7-NYRsTw.

FlagEvalMM: A Flexible Framework for Comprehensive Multimodal Model Evaluation

Zheqi He, Yesheng Liu, Jing-Shu Zheng, Xuejing Li, Jin-Ge Yao, Bowen Qin, richeng xuan, Xi Yang

We present FlagEvalMM, an open-source evaluation framework designed to comprehensively assess multimodal models across a diverse range of vision-language understanding and generation tasks, such as visual question answering, text-to-image/video generation, and image-text retrieval. We decouple model inference from evaluation through an independent evaluation service, thus enabling flexible resource allocation and seamless integration of new tasks and models. Moreover, FlagEvalMM utilizes advanced inference acceleration tools (e.g., vLLM, SGLang) and asynchronous data loading to significantly enhance evaluation efficiency. Extensive experiments show that FlagEvalMM offers accurate and efficient insights into model strengths and limitations, making it a valuable tool for advancing multimodal research. The framework is publicly accessible at https://github.com/flageval-baai/FlagEvalMM, with a demonstration video available at https://youtu.be/L7EtacjoM0k.

Unifying Language Agent Algorithms with Graph-based Orchestration Engine for Reproducible Agent Research

Qianqian Zhang, Jiajia Liao, Heting Ying, Yibo Ma, Haozhan Shen, jingcheng li, Peng Liu, Lu Zhang, Chunxin Fang, Kyusong Lee, Ruochen Xu, Tiancheng Zhao

Language agents powered by large language models (LLMs) have demonstrated remarkable capabilities in understanding, reasoning, and executing complex tasks. However, developing robust agents presents significant challenges: substantial engineering overhead, lack of standardized components, and insufficient evaluation frameworks for fair comparison. We introduce Agent Graph-based Orchestration for Reasoning and Assessment (AGORA), a flexible and extensible framework that addresses these challenges through three key contributions: (1) a modular architecture with a graph-based workflow engine, efficient memory management, and clean component abstraction; (2) a comprehensive suite of reusable agent algorithms implementing state-of-the-art reasoning approaches; and (3) a rigorous evaluation framework enabling systematic comparison across multiple dimensions. Through extensive experiments on mathematical reasoning and multimodal tasks, we evaluate various agent algorithms across different LLMs, revealing important insights about their relative strengths and applicability. Our results demonstrate that while sophisticated reasoning approaches can enhance agent capabilities, simpler methods like Chain-of-Thought often exhibit robust performance with significantly lower computational overhead. AGORA not only simplifies language agent development but also establishes a foundation for reproducible agent research through standardized evaluation protocols. We made a demo video at: https://www.youtube.com/watch?v=WRH-F1zegKI. The comparison agent of algorithms is also available at https://huggingface.co/spaces/omlab/open-agent-leaderboard. Source code of AGORA can be found at https://github.com/om-ai-lab/OmAgent.

Dia-Lingle: A Gamified Interface for Dialectal Data Collection

Jiugeng Sun, Rita Sevastjanova, Sina Ahmadi, Rico Sennrich, Mennatallah El-Assady

Dialects suffer from the scarcity of computational textual resources as they exist predominantly in spoken rather than written form and exhibit remarkable geographical diversity. Collecting dialect data and subsequently integrating it into current language technologies present significant obstacles. Gamification has been proven to facilitate remote data collection processes with great ease and on a substantially wider scale. This paper introduces Dia-Lingle, a gamified interface aimed to improve and facilitate dialectal data collection tasks such as corpus expansion and dialect labelling. The platform features two key components: the first challenges users to rewrite sentences in their dialects, identifies them through a classifier and solicits feedback, and the other one asks users to match sentences to their geographical locations. Dia-Lingle combines active learning with gamified difficulty levels, strategically encouraging prolonged user engagement while efficiently enriching the dialect corpus. Usability evaluation shows that our interface demonstrates high levels of user satisfaction. We provide the link to Dia-Lingle: https://dia-lingle.ivia.ch/, and demo video: https://youtu.be/0QyJsB8ym64.

AutoAlign: Get Your LLM Aligned with Minimal Annotations

Xinyu Lu, Dong Xu, Chunkang Zhang, Xinyan Guan, Junxiang Wang, Qingyu Zhang, Pengbo Wang, Yingzhi Mao, Hao Xiang, Xueru Wen, Zichao Li, Yaojie Lu, Hongyu Lin, Le Sun, Xianpei Han

Automated Alignment refers to a set of algorithms designed to align Large Language Models (LLMs) with human intentions and values while minimizing manual intervention. However, it faces challenges such as algorithmic diversity and excessively convoluted workflows. We present AutoAlign, an open-source toolkit that offers: (1) a unified framework integrating mainstream automated algorithms through a consistent interface, and (2) an accessible workflow supporting one-click execution for prompt synthesis, automatic alignment signal construction, and iterative model training. Our toolkit enables easy reproduction of existing results through extensive benchmarks and facilitates the development of novel approaches via modular components. It includes implementations for both highly efficient inference and training, as well as low-resource training. By standardizing automated alignment methodologies and providing accessible implementations, AutoAlign lowers the barriers to building customized aligned models and supports academic research.

Efficient Annotator Reliablity Assessment with EffiARA

Owen Cook, Jake A Vasilakes, Ian Roberts, Xingyi Song

Data annotation is an essential component of the machine learning pipeline; it is also a costly and time-consuming process. With the introduction of transformer-based models, annotation at the document level is increasingly popular; however, there is no standard framework for structuring such tasks. The EffiARA annotation framework is, to our knowledge, the first project to support the whole annotation pipeline, from understanding the resources required for an annotation task to compiling the annotated dataset and gaining insights into the reliability of individual annotators as well as the dataset as a whole. The framework's efficacy is supported by two previous studies: one improving classification performance through annotator-reliability-based soft label aggregation and sample weighting, and the other increasing the overall agreement among annotators through removing identifying and replacing an unreliable annotator. This work introduces the EffiARA Python package and its accompanying webtool, which provides an accessible graphical user interface for the system. We open-source the EffiARA Python package at https://github.com/MiniEggz/EffiARA and the webtool is publicly accessible at https://effiara.gate.ac.uk.

Value Compass Leaderboard: A Comprehensive, Generative and Self-Evolving Platform for LLMs' Value Evaluation

Jing Yao, Xiaoyuan Yi, Shitong Duan, Jindong Wang, Yuzhuo Bai, Muhua Huang, Yang Ou, Scarlett Li, Peng Zhang, Tun Lu, Zhicheng Dou, Maosong Sun, James Evans, Xing Xie

As large language models (LLMs) are gradually integrated into human life, assessing their underlying values becomes essential for understanding their risks and alignment with specific cultural preferences. Despite growing efforts, current value evaluation methods face two key challenges. C1. Evaluation Validity: Static benchmarks fail to reflect intended values or yield informative results due to data contamination or ceiling effect. C2. Result Interpretation: They typically reduce the pluralistic and often incommensurable values to one-dimensional scores, which hinders users from gaining meaningful insights and guidance. To address these challenges, we present Value Compass Leaderboard, the first interactive platform specially devised for comprehensive value diagnosis of LLMs. It (1) grounds evaluations in multiple basic value systems from social science; (2) develops a generative evolving evaluation schema that automatically and dynamically creates real-world test items that co-evolve with ever-advancing LLMs; (3) offers multi-faceted result interpretation, including: (i) fine-grained scores and case studies across 27 value dimensions for 33 leading LLMs, (ii) customized comparisons, and (iii) visualized analysis of LLMs' alignment with cultural values. We hope Value Compass Leaderboard. serves as a navigator for further enhancing LLMs' safety and alignment, further benefiting their responsible and adaptive development.

Session 7: Semantics: Lexical and Sentence-Level (Tuesday 29 July, 10:30-12:00, Hall 5X)

Textagon: Boosting Language Models with Theory-guided Parallel Representations

John P. Lalor, Ruiyang Qin, David Dobolyi, Ahmed Abbasi

Pretrained language models have significantly advanced the state of the art in generating distributed representations of text. However, they do not account for the wide variety of available expert-generated language resources and lexicons that explicitly encode linguistic/domain knowledge. Such lexicons can be paired with learned embeddings to further enhance NLP prediction and linguistic inquiry. In this work we present Textagon, a Python package for generating parallel representations for text based on predefined lexicons and selecting representations that provide the most information. We discuss the motivation behind the software, its implementation, as well as two case studies for its use to demonstrate operational utility.

LiDARR: Linking Document AMRs with Referents Resolvers

Jon Cai, Kristin Wright-Bettner, Zekun Zhao, Shafiuddin Rehan Ahmed, Abijith Trichur Ramachandran, Jeffrey Flanigan, Martha Palmer, James Martin

In this paper, we present LiDARR (**Li**nking **D**ocument **A**MRs with **R**eferents **R**esolvers), a web tool for semantic annotation at the document level using the formalism of Abstract Meaning Representation (AMR). LiDARR streamlines the creation of comprehensive knowledge graphs from natural language documents through semantic annotation. The tool features a visualization and interactive user interface, transforming document-level AMR annotation into an models-facilitated verification process. This is achieved through the integration of an AMR-to-surface alignment model and a coreference resolution model. Additionally, we incorporate PropBank rolesets into LiDARR to extend implicit roles in annotated AMR, allowing implicit roles to be linked through the coreference chains via AMRs.

Session 7: Speech recognition, text-to-speech and spoken language understanding (Tuesday 29 July, 10:30-12:00, Hall 4/5)

MMS-LLaMA: Efficient LLM-based Audio-Visual Speech Recognition with Minimal Multimodal Speech Tokens

Jeong Hun Yeo, Hyeongseop Rha, Se Jin Park, Yong Man Ro

Audio-Visual Speech Recognition (AVSR) achieves robust speech recognition in noisy environments by combining auditory and visual information. However, recent Large Language Model (LLM) based AVSR systems incur high computational costs due to the high temporal resolution of audio-visual speech processed by LLMs. In this work, we introduce an efficient multimodal speech LLM framework that minimizes token length while preserving essential linguistic content. Our approach employs an early av-fusion module for streamlined feature integration, an audio-visual speech Q-Former that dynamically allocates tokens based on input duration, and a refined query allocation strategy with a speech rate predictor to adjust token allocation according to speaking speed of each audio sample. Extensive experiments on the LRS3 dataset show that our method achieves state-of-the-art performance with a WER of 0.74% while using only 2.8 tokens per second. Moreover, our approach not only reduces token usage by 88.8% compared to the previous multimodal speech LLM framework, but also improves computational efficiency by reducing FLOPs by 36%.

Session 7: Student Research Workshop (Tuesday 29 July, 10:30-12:00, Hall 4/5)

Beyond the Gold Standard in Analytic Automated Essay Scoring

Gabrielle Gaudeau

Originally developed to reduce the manual burden of grading standardised language tests, Automated Essay Scoring (AES) research has long focused on holistic scoring methods which offer minimal formative feedback in the classroom. With the increasing demand for technological tools that support language acquisition, the field is turning to analytic AES (evaluating essays according to different linguistic traits). This approach holds promise for generating more detailed essay feedback, but relies on analytic scoring data that is both more cognitively demanding for humans to produce, and prone to bias. The dominant paradigm in AES is to aggregate disagreements between raters into a single gold-standard label, which fails to account for genuine examiner variability. In an attempt to make AES more representative and trustworthy, we propose to explore the sources of disagreements and lay out a novel AES system design that learns from individual raters instead of the gold standard labels.

Confidence and Stability of Global and Pairwise Scores in NLP Evaluation

Georgii Levtsov, Dmitry Ustalov

With the advent of highly capable instruction-tuned neural language models, benchmarking in natural language processing (NLP) is increasingly shifting towards pairwise comparison leaderboards, such as LMSYS Arena, from traditional global pointwise scores (e.g., GLUE, BIG-bench, SWE-bench). This paper empirically investigates the strengths and weaknesses of both global scores and pairwise comparisons to aid decision-making in selecting appropriate model evaluation strategies. Through computational experiments on synthetic and real-world datasets using standard global metrics and the popular Bradley–Terry model for pairwise comparisons, we found that while global scores provide more reliable overall rankings, they can underestimate strong models with rare, significant errors or low confidence. Conversely, pairwise comparisons are particularly effective for identifying strong contenders among models with lower global scores, especially where quality metrics are hard to define (e.g., text generation), though they require more comparisons to converge if ties are frequent.

Zero-shot prompt-based classification: topic labeling in times of foundation models in German Tweets

Simon Mnker, Kai Kugler, Achim Rettinger

Filtering and annotating textual data are routine tasks in many areas, like social media or news analytics. Automating these tasks allows to scale the analyses wrt. speed and breadth of content covered and decreases the manual effort required. Due to technical advancements in Natural Language Processing, specifically the success of large foundation models, a new tool for automating such annotation processes by using a text-to-text interface given written guidelines without providing training samples has become available. In this work, we assess these advancements \textit{in-the-wild} by empirically testing them in an annotation task on German Twitter data about social and political European crises. We compare the prompt-based results with our human annotation and preceding classification approaches, including Naive Bayes and a BERT-based fine-tuning/domain adaptation pipeline. Our results show that the prompt-based approach – despite being limited by local computation resources during the model selection – is comparable with the fine-tuned BERT but without any annotated training data. Our findings emphasize the ongoing paradigm shift in the NLP landscape, i.e., the unification of downstream tasks and elimination of the need for pre-labeled training data.

HYPEROFA: Expanding LLM Vocabulary to New Languages via Hypernetwork-Based Embedding Initialization

Enes zeren, Yihong Liu, Hinrich Schuetze

Many pre-trained language models (PLMs) exhibit suboptimal performance on mid- and low-resource languages, largely due to limited exposure to these languages during pre-training. A common strategy to address this is to introduce new tokens specific to the target languages, initialize their embeddings, and apply continual pre-training on target-language data. Among such methods, OFA (Liu et al., 2024a) proposes a similarity-based subword embedding initialization heuristic that is both effective and efficient. However, OFA restricts target-language token embeddings to be convex combinations of a fixed number of source-language embeddings, which may limit expressiveness. To overcome this limitation, we propose HYPEROFA, a hypernetwork-based approach for more adaptive token embedding initialization. The hypernetwork is trained to map from an external multilingual word vector space to the PLM's token embedding space using source-language tokens. Once trained, it can generate flexible embeddings for target-language tokens, serving as a good starting point for continual pretraining. Experiments demonstrate that HYPEROFA consistently outperforms random initialization baseline and matches or exceeds the performance of OFA in both continual pre-training convergence and downstream task performance. We make the code publicly available.

SEPSIS: I Can Catch Your Lies -- A New Paradigm for Deception Detection

Anku Rani, Dwip Dalal, Shreya Gautam, Pankaj Gupta, Vinija Jain, Aman Chadha, Amit Sheth, Amitava Das

Deception is the intentional practice of twisting information. It is a nuanced societal practice deeply intertwined with human societal evolution, characterized by a multitude of facets. This research explores the problem of deception through the lens of psychology, employing a framework that categorizes deception into three forms: lies of omission, lies of commission, and lies of influence. The primary focus of this study is specifically on investigating only lies of omission. We propose a novel framework for deception detection leveraging NLP techniques. We curated an annotated dataset of 876,784 samples by amalgamating a popular large-scale fake news dataset and scraped news headlines from the Twitter handle of "Times of India", a well-known Indian news media house. Each sample has been labeled with four layers, namely: (i) the type of omission (speculation, bias, distortion, sounds factual, and opinion), (ii) colors of lies (black, white, grey, and red), and (iii) the intention of such lies (to influence, gain social prestige, etc) (iv) topic of lies (political, educational, religious, racial, and ethnicity). We present a novel multi-task learning [MTL] pipeline that leverages the dataless merging of fine-tuned language models to address the deception detection task mentioned earlier. Our proposed model achieved an impressive F1 score of 0.87, demonstrating strong performance across all layers including the type, color, intent, and topic aspects of deceptive content. Finally, our research aims to explore the relationship between the lies of omission and propaganda techniques. To accomplish this, we conducted an in-depth analysis, uncovering compelling findings. For instance, our analysis revealed a significant correlation between loaded language and opinion, shedding light on their interconnectedness. To encourage further research in this field, we will publicly release the SEPSIS dataset and models.

Grouped Sequency-arranged Rotation: Optimizing Rotation Transformation for Quantization for Free

Euntae Choi, Sumin Song, Woosang Lim, Sungjoo Yoo

Large Language Models (LLMs) face deployment challenges due to high computational costs, and while Post-Training Quantization (PTQ) offers a solution, existing rotation-based methods struggle at very low bit-widths like 2-bit. We introduce a novel, training-free approach to construct an improved rotation matrix, addressing the limitations of current methods. The key contributions include leveraging the Walsh-Hadamard transform with sequency ordering, which clusters similar frequency components to reduce quantization error compared to standard Hadamard matrices, significantly improving performance. Furthermore, we propose a Grouped Sequency-arranged Rotation (GSR) using block-diagonal matrices with smaller Walsh blocks, effectively isolating outlier impacts and achieving performance comparable to optimization-based methods without requiring any training. Our method demonstrates robust performance on reasoning tasks and Perplexity (PPL) score on WikiText-2. Our method also enhances results even when applied over existing learned rotation techniques.

Neuron-Level Language Tag Injection Improves Zero-Shot Translation Performance

Jay Royal Orten, Ammon Shurtz, Nancy Fulda, Stephen D. Richardson

Language tagging, a method whereby source and target inputs are prefixed with a unique language token, has become the de facto standard for conditioning Multilingual Neural Machine Translation (MNMT) models on specific language directions. This conditioning can manifest effective zero-shot translation abilities in MT models at scale for many languages. Expanding on previous work, we propose a novel method of language tagging for MNMT, injection, in which the embedded representation of a language token is concatenated to the input of every linear layer. We explore a variety of different tagging methods, with and without injection, showing that injection improves zero-shot translation performance with up to a 2+ BLEU score point gain for certain language directions in our dataset.

Your Pretrained Model Tells the Difficulty Itself: A Self-Adaptive Curriculum Learning Paradigm for Natural Language Understanding

Qi Feng, Yihong Liu, Hinrich Schuetze

Curriculum learning is a widely adopted training strategy in natural language processing (NLP), where models are exposed to examples organized by increasing difficulty to enhance learning efficiency and performance. However, most existing approaches rely on manually defined difficulty metrics -- such as text length -- which may not accurately reflect the model's own perspective. To overcome this limitation, we present a self-adaptive curriculum learning paradigm that prioritizes fine-tuning examples based on difficulty scores predicted by pre-trained language models (PLMs) themselves. Building on these scores, we explore various training strategies that differ in the ordering of examples for the fine-tuning: from easy-to-hard, hard-to-easy, to mixed sampling. We evaluate our method on four natural language understanding (NLU) datasets covering both binary and multi-class classification tasks. Experimental results show that our approach leads to faster convergence and improved performance compared to standard random sampling.

iPrOp: Interactive Prompt Optimization for Large Language Models with a Human in the Loop

Jiahui Li, Roman Klinger

Prompt engineering has made significant contributions to the era of large language models, yet its effectiveness depends on the skills of a prompt author. This paper introduces \textit{iPrOp}, a novel interactive prompt optimization approach, to bridge manual prompt engineering and automatic prompt optimization while offering users the flexibility to assess evolving prompts. We aim to provide users with task-specific guidance to enhance human engagement in the optimization process, which is structured through prompt variations, informative instances, predictions generated by large language models along with their corresponding explanations, and relevant performance metrics. This approach empowers users to choose and further refine the prompts based on their individual preferences and needs. It can not only assist non-technical domain experts in generating optimal prompts tailored to their specific tasks or domains, but also enable to study the intrinsic parameters that influence the performance of prompt optimization. The evaluation shows that our approach has the capability to generate improved prompts, leading to enhanced task performance.

Bias Amplification: Large Language Models as Increasingly Biased Media

Ze Wang, Zekun Wu, Jeremy Zhang, Xin Guan, Navya Jain, Skylar Lu, Saloni Gupta, Adriano Koshiyama

Model collapse—a phenomenon characterized by performance degradation due to iterative training on synthetic data—has been widely studied. However, its implications for bias amplification, the progressive intensification of pre-existing societal biases in Large Language Models (LLMs), remain significantly underexplored, despite the growing influence of LLMs in shaping online discourse. In this paper, we introduce a open, generational, and long-context benchmark specifically designed to measure political bias amplification in LLMs, leveraging sentence continuation tasks derived from a comprehensive dataset of U.S. political news. Our empirical study using GPT-2 reveals consistent and substantial political bias intensification (e.g., right-leaning amplification) over iterative synthetic training cycles. We evaluate three mitigation strategies—Overfitting, Preservation, and Accumulation—and demonstrate that bias amplification persists independently of model collapse, even when the latter is effectively controlled. Furthermore, we propose a mechanistic analysis approach that identifies neurons correlated with specific phenomena during inference through regression and statistical tests. This analysis uncovers largely distinct neuron populations driving bias amplification and model collapse, underscoring fundamentally different underlying mechanisms. Finally, we supplement our empirical findings with theoretical intuition that explains the separate origins of these phenomena, guiding targeted strategies for bias mitigation.

FaithfulSAE: Towards Capturing Faithful Features with Sparse Autoencoders without External Datasets Dependency

Seonglae Cho, Harryn Oh, Donghyun Lee, Luis Eduardo Rodrigues Vieira, Andrew Bermingham, Ziad El Sayed

Sparse Autoencoders (SAEs) have emerged as a promising solution for decomposing large language model representations into interpretable features. However, Paulo \& Belrose (2025) have highlighted instability across different initialization seeds, and Heap et al. (2025) have pointed out that SAEs may not capture model-internal features. These problems likely stem from training SAEs on external datasets—either collected from the Web or generated by another model—which may contain out-of-distribution (OOD) data beyond the model's generalisation capabilities. This can result in hallucinated SAE features, which we term "Fake Features", that misrepresent the model's internal activations. To address these issues, we propose FaithfulSAE, a method that trains SAEs on the model's own synthetic dataset. Using FaithfulSAEs, we demonstrate that training SAEs on less-OOD instruction datasets results in SAEs being more stable across seeds. Notably, FaithfulSAEs outperform SAEs trained on webbased datasets in the SAE probing task and exhibit a lower Fake Feature Ratio in 5 out of 7 models. Overall, our approach eliminates the dependency on external datasets, advancing interpretability by better capturing model-internal features while highlighting the often neglected importance of SAE training datasets.

Translating Movie Subtitles by Large Language Models using Movie-meta Information

Ashmari Pramodya, Yusuke Sakai, Justin Vasselli, Hidetaka Kamigaito, Taro Watanabe

Large language models (LLMs) have advanced natural language processing by understanding, generating, and manipulating texts. Although recent studies have shown that prompt engineering can reduce computational effort and potentially improve translation quality, prompt designs specific to different domains remain challenging. Besides, movie subtitle translation is particularly challenging and understudied, as it involves handling colloquial language, preserving cultural nuances, and requires contextual information such as the movie's theme and storyline to ensure accurate meaning. This study aims to fill this gap by focusing on the translation of movie subtitles through the use of prompting strategies that incorporate the movie's meta-information, e.g., movie title, summary, and genre. We build a multilingual dataset which aligns the OpenSubtitles dataset with their corresponding Wikipedia articles and investigate different prompts and their effect on translation performance. Our experiments with GPT-3.5, GPT-4o, and LLaMA-3 models have shown that the presence of meta-information improves translation accuracy. These findings further emphasize the importance of designing appropriate prompts and highlight the potential of LLMs to enhance subtitle translation quality.

Small Models, Big Impact: Efficient Corpus and Graph-Based Adaptation of Small Multilingual Language Models for Low-Resource Languages

Daniil Gurgurov, Ivan Vykopal, Josef van Genabith, Simon Ostermann

Low-resource languages (LRLs) face significant challenges in natural language processing (NLP) due to limited data. While current state-of-the-art large language models (LLMs) still struggle with LRLs, smaller multilingual models (mLMs) such as mBERT and XLM-R offer greater promise due to a better fit of their capacity to low training data sizes. This study systematically investigates parameter-efficient adapter-based methods for adapting mLMs to LRLs, evaluating three architectures: Sequential Bottleneck, Invertible Bottleneck, and Low-Rank Adaptation. Using unstructured text from GlotCC and structured knowledge from ConceptNet, we show that small adaptation datasets (e.g., up to 1 GB of free-text or a few MB of knowledge graph data) yield gains in intrinsic (masked language modeling) and extrinsic tasks (topic classification, sentiment analysis, and named entity recognition). We find that Sequential Bottleneck adapters excel in language modeling, while Invertible Bottleneck adapters slightly outperform other methods on downstream tasks due to better embedding alignment and larger parameter counts. Adapter-based methods match or outperform full fine-tuning while using far fewer parameters, and smaller mLMs prove more effective for LRLs than massive LLMs like LLaMA-3, GPT-4, and DeepSeek-R1-based distilled models. While adaptation improves performance, pre-training data size remains the dominant factor, especially for languages with extensive pre-training coverage.

Exploring the Effect of Nominal Compound Structure in Scientific Texts on Reading Times of Experts and Novices

Isabell Landwehr, Marie-Pauline Krielke, Stefania Degaetano-Ortlieb

We explore how different types of nominal compound complexity in scientific writing, in particular different types of compound structure, affect the reading times of experts and novices. We consider both in-domain and out-of-domain reading and use PoTeC (Jakobi et al. 2024), a corpus containing eye-tracking data of German native speakers reading passages from scientific textbooks. Our results suggest that some compound types are associated with longer reading times and that experts may not only have an advantage while reading in-domain texts, but also while reading out-of-domain.

From Ambiguity to Accuracy: The Transformative Effect of Coreference Resolution on RAG systems

Youngjoon Jang, Seongtae Hong, Junyoung Son, Sungjin Park, Chanjun Park, Heuiseok Lim

Retrieval-Augmented Generation (RAG) has emerged as a crucial framework in natural language processing (NLP), improving factual consistency and reducing hallucinations by integrating external document retrieval with large language models (LLMs). However, the effectiveness of RAG is often hindered by coreferential complexity in retrieved documents, which can introduce ambiguity and interfere with in-context learning. In this study, we systematically investigate how entity coreference affects both document retrieval and generative performance in RAG-based systems, focusing on retrieval relevance, contextual understanding, and overall response quality. We demonstrate that coreference resolution enhances retrieval effectiveness and improves question-answering (QA) performance. Through comparative analysis of different pooling strategies in retrieval tasks, we find that mean pooling demonstrates superior context capturing ability after applying coreference resolution. In QA tasks, we discover that smaller models show greater improvement from the disambiguation process, likely due to their limited inherent capacity for handling referential ambiguity. With these findings, this study aims to provide a deeper understanding of the challenges posed by coreferential complexity in RAG, offering guidance for improving retrieval and generation in knowledge-intensive AI applications.

Quantifying the Influence of Irrelevant Contexts on Political Opinions Produced by LLMs

Samuele D'Avenia, Valerio Basile

Several recent works have examined the generations produced by large language models (LLMs) on subjective topics such as political opinions and attitudinal questionnaires. There is growing interest in controlling these outputs to align with specific users or perspectives using model steering techniques. However, several studies have highlighted unintended and unexpected steering effects, where minor changes in the prompt or irrelevant contextual cues influence model-generated opinions. This work empirically tests how irrelevant information can systematically bias model opinions in specific directions. Using the Political Compass Test questionnaire, we conduct a detailed statistical analysis to quantify these shifts using the opinions generated by LLMs in an open-generation setting. The results demonstrate that even seemingly unrelated contexts consistently alter model responses in predictable ways, further highlighting challenges in ensuring the robustness and reliability of LLMs when generating opinions on subjective topics.

Making Sense of Korean Sentences: A Comprehensive Evaluation of LLMs through KoSEnd Dataset

Seunguk Yu, Kyeonghyun Kim, JungMin Yun, YoungBin Kim

Although LLMs have made significant progress in handling various languages, there are still concerns about their effectiveness with low-resource agglutinative languages compared to languages such as English. In this study, we focused on Korean, a language known for its complex sentence endings, and evaluated LLMs on this challenging aspect. We introduce the Korean Sentence Endings (KoSEnd) dataset, which includes 3,000 sentences and 45,000 sentence ending labels. These were collected from diverse sources to cover a wide range of contexts. We evaluated 11 models to assess their understanding of Korean sentence endings, analyzing them based on parameter count and prediction consistency. Notably, we observed that informing models about the possibility of missing sentence endings led to improved performance, demonstrating the influence of explicitly considering certain linguistic features.

Enhancing Software Requirements Engineering with Language Models and Prompting Techniques: Insights from the Current Research and Future Directions

Moemen Ebrahim, Shawkat Guirguis, Christine Basta

Large Language Models (LLMs) offer transformative potential for Software Requirements Engineering (SRE), yet critical challenges, including domain ignorance, hallucinations, and high computational costs, hinder their adoption. This paper proposes a conceptual framework that integrates Small Language Models (SLMs) and Knowledge-Augmented LMs (KALMs) with LangChain to address these limitations systematically. Our approach combines: (1) SLMs for efficient, locally deployable requirements processing, (2) KALMs enhanced with Retrieval-Augmented Generation (RAG) to mitigate domain-specific gaps, and (3) LangChain for structured, secure workflow orchestration. We identify and categorize six technical challenges and two research gaps through a systematic review of LLM applications in SRE. To guide practitioners, we distill evidence-based prompt engineering guidelines (Context, Language, Examples, Keywords) and propose prompting strategies (e.g., Chain-of-Verification) to improve output reliability. The paper establishes a theoretical foundation for scalable, trustworthy AI-assisted SRE and outlines future directions, including domain-specific prompt templates and hybrid validation pipelines.

Question Decomposition for Retrieval-Augmented Generation

Paul J. L. Ammann, Jonas Golde, Alan Akbik

Grounding large language models (LLMs) in verifiable external sources is a well-established strategy for generating reliable answers. Retrieval-augmented generation (RAG) is one such approach, particularly effective for tasks like question answering: it retrieves passages that are semantically related to the question and then conditions the model on this evidence. However, multi-hop questions, such as \textit{``Which company among NVIDIA, Apple, and Google made the biggest profit in 2023?,"} challenge RAG because relevant facts are often distributed across multiple documents rather than co-occurring in one source, making it difficult for standard RAG to retrieve sufficient information. To address this, we propose a RAG pipeline that incorporates question decomposition: (i) an LLM decomposes the original query into sub-questions, (ii) passages are retrieved for each sub-question, and (iii) the merged candidate pool is reranked to improve the coverage and precision of the retrieved evidence. We show that question decomposition effectively assembles complementary documents, while reranking reduces noise and promotes the most relevant passages before answer generation. We evaluate our approach on the MultiHop-RAG and HotpotQA, showing gains in retrieval (MRR@10: +36.7\%) and answer accuracy (F1: +11.6\%) over standard RAG baselines. The pipeline is a practical, drop-in enhancement requiring no task-specific training or specialized indexing.

Spanish Dialect Classification: A Comparative Study of Linguistically Tailored Features, Unigrams and BERT Embeddings

Laura Zeidler, Chris Jenkins, Filip Mileti, Sabine Schulte im Walde

The task of automatic dialect classification is typically tackled using traditional machine-learning models with bag-of-words unigram features. We explore two alternative methods for distinguishing dialects across 20 Spanish-speaking countries: (i) Support vector machine and decision tree models were trained on dialectal features tailored to the Spanish dialects, combined with standard unigrams. (ii) A pre-trained BERT model was fine-tuned on the task. Results show that the tailored features generally did not have a positive impact on traditional model performance, but provide a salient way of representing dialects in a content-agnostic manner. The BERT model wins over traditional models but with only a tiny margin, while sacrificing explainability and interpretability.

MA-COIR: Leveraging Semantic Search Index and Generative Models for Ontology-Driven Biomedical Concept Recognition

Shanshan liu, Noriki Nishida, Rumana Ferdous Munne, Narumi Tokunaga, Yuki Yamagata, Kouji Kozaki, Yuji Matsumoto

Recognizing biomedical concepts in the text is vital for ontology refinement, knowledge graph construction, and concept relationship discovery. However, traditional concept recognition methods, relying on explicit mention identification, often fail to capture complex concepts not explicitly stated in the text. To overcome this limitation, we introduce MA-COIR, a framework that reformulates concept recognition as an indexing-recognition task. By assigning semantic search indexes (ssIDs) to concepts, MA-COIR resolves ambiguities in ontology entries and enhances recognition efficiency. Using a pretrained BART-based model fine-tuned on small datasets, our approach reduces computational requirements to facilitate adoption by domain experts. Furthermore, we incorporate large language model (LLM)-generated queries and synthetic data to improve recognition in low-resource settings. Experimental results on three scenarios (CDR, HPO, and HOIP) highlight the effectiveness of MA-COIR in recognizing both explicit and implicit concepts without the need for mention-level annotations during inference, advancing ontology-driven concept recognition in biomedical domain applications. Our code and constructed data are available at https://github.com/sl-633/macoir-master.

LibVulnWatch: A Deep Assessment Agent System and Leaderboard for Uncovering Hidden Vulnerabilities in Open-Source AI Libraries

Zekun Wu, Seonglae Cho, Umar Mohammed, CRISTIAN ENRIQUE MUNOZ VILLALOBOS, Kleyton Da Costa, Xin Guan, Theo King, Ze Wang, Emre Kazim, Adriano Koshiyama

Open-source AI libraries are foundational to modern AI systems, yet they present significant, underexamined risks spanning security, licensing, maintenance, supply chain integrity, and regulatory compliance. We introduce LibVulnWatch, a system that leverages recent advances in large language models and agentic workflows to perform deep, evidence-based evaluations of these libraries. Built on a graph-based orchestration of specialized agents, the framework extracts, verifies, and quantifies risk using information from repositories, documentation, and vulnerability databases. LibVulnWatch produces reproducible, governance-aligned scores across five critical domains, publishing results to a public leaderboard for ongoing ecosystem monitoring. Applied to 20 widely used libraries—including ML frameworks, LLM inference engines, and agent orchestration tools—our approach covers up to 88\% of OpenSSF Scorecard checks while surfacing up to 19 additional risks per library, such as critical RCE vulnerabilities, missing SBOMs, and regulatory gaps. By integrating advanced language technologies with the practical demands of software risk assessment, this work demonstrates a scalable, transparent mechanism for continuous supply chain evaluation and informed library selection.

Interactive Text Games: Lookahead Is All You Need!

Hosein Rezaei, James Alfred Walker, Frank Soboczenski

The cross-modal grounding of LLMs has recently garnered significant attention, while grounding them in textual interactions has been less explored. As the first of its kind, the GLAM framework utilises LLMs as agents in interactive text-based games to investigate their grounding capabilities. However, it faces the challenge of low computational efficiency, which hinders further experiments. This paper proposes the use of Lookahead models for action selection, demonstrating through empirical results that the approach can substantially improve training speed, achieving performance gains relative to the size of the action space.

The Evolution of Gen Alpha Slang: Linguistic Patterns and AI Translation Challenges

Ishita, Radhika Mamidi

Generation Alpha (born 2010-2024) is the first generation fully raised within the digital ecosystem. They exhibit unique linguistic behaviours influenced by rampant online communication and platform-specific cultures. This study examines the rapid evolution of Gen Alpha slang through a comparative analysis of Millennial and Gen Z vernacular. We identify three core linguistic patterns: extreme lexical compression, digital culture-driven semantic shifts and part-of-speech conversion. We construct a comprehensive slang corpus sourced from online platforms and evaluate the performance of four AI translation systems (viz. Google Translate, ChatGPT 4, Gemini 1.0, DeepSeek v3) on over 100 slang terms. Our results reveal significant translation challenges rooted in culturally-bound terms from gaming, meme culture, and mental health discourse. Most errors are the result of inadequate cultural contextualization, with literal translations dominating the error patterns. Our findings highlight the critical limitations in current language models and emphasize the need for adaptive, culturally sensitive and context-aware frameworks that can handle the dynamic lexicon of evolving youth vernacular.

Light-Weight Hallucination Detection using Contrastive Learning for Conditional Text Generation

Miyu Yamada, Yuki Arase

We propose a simple and light-weight, yet effective hallucination detection method for conditional text generation. Hallucinated outputs include information that is either absent from and/or difficult to infer from the input context. Leveraging this feature, we add contrastive learning to the hallucination detection classifier to pull faithful outputs and input contexts together while pushing hallucinated outputs apart. Experimental results confirm that our method on top of RoBERTa improves binary hallucination detection performance, outperforming much larger GPT-4o prompting. Remarkably, our method shows higher performance for outputs where hallucinated spans are sparse.

Cross-Genre Learning for Old English Poetry POS Tagging

Irene Miani, Sara Stymne, Gregory Darwin

Poetry has always distinguished itself from other literary genres in many ways, including grammatically and syntactically. These differences are evident not only in modern literature but also in earlier stages. Linguistic analysis tools struggle to address these differences. This paper focuses on the dichotomy between Old English poetry and prose, specifically in the context of the POS tagging task. Two annotated corpora representing each genre were analyzed to show that there are several types of structural differences between Old English poetry and prose. For POS tagging, we conduct experiments on both a detailed tag set with over 200 tags and a mapping to the UPOS tag set with 17 tags. We establish a baseline and conduct two cross-genre experiments to investigate the effect of different proportions of prose and poetry data. Across both tag sets, our results indicate that if the divergence between two genres is substantial, simply increasing the quantity of training data from the support genre does not necessarily improve prediction accuracy. However, incorporating even a small amount of target data can lead to better performance compared to excluding it entirely. This study not only highlights the linguistic differences between Old English poetry and prose but also emphasizes the importance of developing effective NLP tools for underrepresented historical languages across all genres.

A Computational Framework to Identify Self-Aspects in Text

Jaya Caporusso, Matthew Purver, Senja Pollak

This Ph.D. proposal introduces a plan to develop a computational framework to identify Self-aspects in text. The Self is a multifaceted construct and it is reflected in language. While it is described across disciplines like cognitive science and phenomenology, it remains underexplored in natural language processing (NLP). Many of the aspects of the Self align with psychological and other well-researched phenomena (e.g., those related to mental health), highlighting the need for systematic NLP-based analysis. In line with this, we plan to introduce an ontology of Self-aspects and a gold-standard annotated dataset. Using this foundation, we will develop and evaluate conventional discriminative models, generative large language models, and embedding-based retrieval approaches against four main criteria: interpretability, ground-truth adherence, accuracy, and computational efficiency. Top-performing models will be applied in case studies in mental health and empirical phenomenology.

Can a Large Language Model Keep My Secrets? A Study on LLM-Controlled Agents

Niklas Hemken, Sai Koneru, Florian Jacob, Hannes Hartenstein, Jan Niehues

Agents controlled by large language models (LLMs) have the potential to assist humans in natural language tasks across various domains and applications, if they are provided with access to confidential data of their principal. When such digital assistants interact with their potentially adversarial environment, confidentiality of the data is at stake. Given a natural language request, we investigate whether an LLM-controlled agent can control access to internal data by considering confidentiality in its response, in a manner similar to humans. For evaluation, we created a synthetic dataset consisting of confidentiality-aware planning and deduction tasks in an organizational access control setting. The dataset was developed from human input, LLM-generated content, and existing datasets. It includes various everyday scenarios in which access to confidential or private information is requested. We utilize our dataset to evaluate the ability to infer confidentiality-aware behavior in such scenarios by differentiating between legitimate and illegitimate access requests. We compare a prompting-based and a fine-tuning-based approach, to evaluate the performance of Llama 3 and GPT-4o-mini in this domain. In addition, we conducted a user study to establish a baseline for human evaluation performance in these tasks. We find humans reached an accuracy of up to 79%. Prompting techniques, such as chain-of-thought and few-shot prompting, yield promising results, but still fall short of real-world applicability and do not surpass human baseline performance. However, we find that fine-tuning significantly improves the agent's ability to make access decisions up to an accuracy of 98%, making it a promising approach for future confidentiality-aware applications when data is available.

Chart Question Answering from Real-World Analytical Narratives

Maeve Hutchinson, Radu Jianu, Aidan Slingsby, Jo Wood, Pranava Madhyastha

We present a new dataset for chart question answering (CQA) constructed from visualization notebooks. The dataset features real-world, multi-view charts paired with natural language questions grounded in analytical narratives. Unlike prior benchmarks, our data reflects ecologically valid reasoning workflows. Benchmarking state-of-the-art multimodal large language models reveals a significant performance gap, with GPT-4.1 achieving an accuracy of 69.3\%, underscoring the challenges posed by this more authentic CQA setting.

Low-Perplexity LLM-Generated Sequences and Where To Find Them

Arthur Wuhrmann, Andrei Kucharavy, Anastasiia Kucherenko

As Large Language Models (LLMs) become increasingly widespread, understanding how specific training data shapes their outputs is crucial for transparency, accountability, privacy, and fairness. To explore how LLMs recall and replicate learned information, we introduce a systematic approach centered on analyzing low-perplexity sequences—high-probability text spans generated by the model. Our pipeline reliably extracts such long sequences across diverse topics while avoiding degeneration, then traces them back to their sources in the training data. Surprisingly, we find that a substantial portion of these low-perplexity spans cannot be mapped to the corpus. For those that do match, we analyze the types of memorization involved and present the distribution of unique documents contributing to these mappings, highlighting the extent of verbatim recall.

CoLeM: A framework for semantic interpretation of Russian-language tables based on contrastive learning

Kirill Tobola, Nikita Dorodnykh

Tables are extensively utilized to represent and store data, however, they often lack explicit semantics necessary for machine interpretation of their contents. Semantic table interpretation is essential for integrating structured data with knowledge graphs, yet existing methods face challenges with Russian-language tables due to limited labeled data and linguistic peculiarities. This paper introduces a contrastive learning approach to minimize reliance on manual labeling and enhance the accuracy of column annotation for rare semantic types. The proposed method adapts contrastive learning for tabular data through augmentations and employs a distilled multilingual BERT model trained on the unlabeled RWT corpus (comprising 7.4 million columns). The resulting table representations are incorporated into the RuTaBERT pipeline, reducing computational overhead. Experimental results demonstrate a micro-F1 score of 97% and a macro-F1 score of 92%, surpassing several baseline approaches. These findings emphasize the efficiency of the proposed method in addressing data sparsity and handling unique features of the Russian language. The results further confirm that contrastive learning effectively captures semantic similarities among columns without explicit supervision, which is particularly vital for rare data types.

Mitigating Hallucination by Integrating Knowledge Graphs into LLM Inference -- a Systematic Literature Review

Robin Wagner, Emanuel Kitzelmann, Ingo Boersch

Large Language Models (LLMs) demonstrate strong performance on different language tasks, but tend to hallucinate -- generate plausible but factually incorrect outputs. Recently, several approaches to integrate Knowledge Graphs (KGs) into LLM inference were published to reduce hallucinations. This paper presents a systematic literature review (SLR) of such approaches. Following established SLR methodology, we identified relevant work by systematically search in different academic online libraries and applying a selection process. Nine publications were chosen for in-depth analysis. Our synthesis reveals differences and similarities of how the KG is accessed, traversed, and how the context is finally assembled. KG integration can significantly improve LLM performance on benchmark datasets and additionally to mitigate hallucination enhance reasoning capabilities, explainability, and access to domain-specific knowledge. We also point out current limitations and outline directions for future work.

Privacy Ripple Effects from Adding or Removing Personal Information in Language Model Training

Jaydeep Borkar, Matthew Jagielski, Katherine Lee, Niloofar Mireshghallah, David A. Smith, Christopher A. Choquette-Choo

Due to the sensitive nature of personally identifiable information (PII), its owners may have the authority to control its inclusion or request its removal from large-language model (LLM) training. Beyond this, PII may be added or removed from training datasets due to evolving dataset curation techniques, because they were newly scraped for retraining, or because they were included in a new downstream fine-tuning stage. We find that the amount and ease of PII memorization is a dynamic property of a model that evolves throughout training pipelines and depends on commonly altered design choices. We characterize three such novel phenomena: (1) similar-appearing PII seen later in training can elicit memorization of earlier-seen sequences in what we call assisted memorization, and this is a significant factor (in our settings, up to 1/3); (2) adding PII can increase memorization of other PII; and (3) removing PII can lead to other PII being memorized.

I Speak for the rboles: Developing a Dependency Treebank for Spanish L2 and Heritage Speakers

Emiliana Pulido, Robert Pugh, Zoey Liu

We introduce the first set of Universal Dependencies (UD) annotations for Spanish learner writing from the UC Davis COWSL2H corpus. Our annotations include lemmatization, POS tagging, and syntactic dependencies. We adapt the existing UD framework for Spanish L1 to account for learner-specific features such as code-switching and non-canonical syntax. A suite of parsing evaluation experiments shows that parsers trained on learner data together with moderate sizes of Spanish L1 data can yield reasonable performance. Our annotations and parsers will be openly accessible to motivate future development of learner-oriented language technologies.

Session 10: CL Journal (Tuesday 29 July, 16:00-17:30, Hall 4/5)

Graded Suspiciousness of Adversarial Texts to Humans

Shakila Mahjabin Tonni Pedro Faustini Mark Dras

Adversarial examples pose a significant challenge to deep neural networks across both image and text domains, with the intent to degrade model performance through carefully altered inputs. Adversarial texts, however, are distinct from adversarial images due to their requirement for semantic similarity and the discrete nature of the textual contents. This study delves into the concept of human suspiciousness, a quality distinct from the traditional focus on imperceptibility found in image-based adversarial examples, where adversarial changes are often desired to be indistinguishable to the human eye even when placed side by side with originals. Although this is generally not possible with text, textual adversarial content must still often remain undetected or non-suspicious to human readers. Even when the text's purpose is to deceive NLP systems or bypass filters, the text is often expected to be natural to read. In this research, we expand the study of human suspiciousness by analyzing how individuals perceive adversarial texts. We gather and publish a novel dataset of Likert-scale human evaluations on the suspiciousness of adversarial sentences, crafted by four widely used adversarial attack methods and assess their correlation with the human ability to detect machine-generated alterations. Additionally, we develop a regression-based model to predict levels of suspiciousness and establish a baseline for future research in reducing the suspiciousness in adversarial text generation. We also demonstrate how the regressor-generated suspicious scores can be incorporated into adversarial generation methods to produce texts that are less likely to be perceived as computer-generated.

Eliciting and Improving the Causal Reasoning Abilities of Large Language Models with Conditional Statements

Xiao Liu, Da Yin, Chen Zhang, Dongyan Zhao, Yansong Feng

Causal reasoning, the ability to identify cause-and-effect relationships, is crucial in human thinking. Although large language models (LLMs) succeed in many NLP tasks, it is still challenging for them to conduct complex causal reasoning like abductive reasoning and counterfactual reasoning. Complex causal structures are rarely expressed explicitly in the text, which could make learning them challenging for LLMs. Given the fact that programming code may expresscausal relations more often and explicitly with conditional statements like if, we want to explore whether large language models of code (Code-LLMs) acquire better causal reasoning abilities, and whether code prompts better describe the causal structure than text prompts. Our experiments show that compared to general-purpose LLMs like LLAMA-2 and GPT-3, Code- LLMs like CODELLAMA and CODEX are significantly better in causal reasoning. Code prompts not only work well for Code-LLMs, but also help improve the performance of most general purpose LLMs. To understand why code prompts are effective, we intervene on the prompts from different aspects, and discover that the programming structure is crucial in code prompt design, while models are more robust towards format perturbations.We further explore whether exposing models with more code with conditional statements aids in enhancing causal reasoning abilities.

BLiMP-NL: A Corpus of Dutch Minimal Pairs and Acceptability Judgments for Language Model Evaluation

Stefan Frank, Michelle Suijkerbuijk, Zoë Prins, Marianne de Heer Kloots, Willem Zuidema, Stefan Frank

We present a corpus of 8400 Dutch sentence pairs, intended primarily for the grammatical evaluation of language models. Each pair consists of a grammatical sentence and a minimally different ungrammatical sentence. The corpus covers 84 paradigms, classified into 22 syntactic phenomena. Ten sentence pairs of each paradigm were created by hand, while the remaining ninety were generated semi-automatically and manually validated afterwards. Nine of the ten hand-crafted sentences of each paradigm are rated for acceptability by at least 30 participants each, and for the same 9 sentences reading times are recorded per word, through self-paced reading.Â Here, we report on the construction of the dataset, the measured acceptability ratings and reading times, as well as the extent to which a variety of language models can be used to predict both the ground-truth grammaticality and human acceptability ratings.

LGDE: Local Graph-based Dictionary Expansion

Juni Schindler, Sneha Jha, Xixuan Zhang, Kilian Buehling, Annett Heft, Mauricio Barahona

We present Local Graph-based Dictionary Expansion (LGDE), a method for data-driven discovery of the semantic neighbourhood of words using tools from manifold learning and network science. At the heart of LGDE lies the creation of a word similarity graph from the geometry of word embeddings followed by local community detection based on graph diffusion. The diffusion in the local graph manifold allows the exploration of the complex nonlinear geometry of word embeddings to capture word similarities based on paths of semantic association, over and above direct pairwise similarities. Exploiting such semantic neighbourhoods enables the expansion of dictionaries of pre-selected keywords, an important step for tasks in information retrieval, such as database queries and online data collection. We validate LGDE on two user-generated English-language corpora and show that LGDE enriches the list of keywords with improved performance relative to methods based on direct word similarities or co-occurrences. We further demonstrate our method through a real-world use case from communication science, where LGDE is evaluated quantitatively on the expansion of a conspiracy-related dictionary from online data collected and analysed by domain experts. Our empirical results and expert user assessment indicate that LGDE expands the seed dictionary with more useful keywords due to the manifold-learning-based similarity network.

A Novel Methodology for Enhancing Cross-Language and Domain Adaptability in Temporal Expression Normalization

Alejandro Sánchez de Castro, Lourdes Araujo, Juan Martinez-Romo

Accurate temporal expression normalization, the process of assigning a numerical value to a temporal expression, is essential for tasks such as timeline creation and temporal reasoning. While rule-based normalization systems are limited in adaptability across different domains and languages, deep-learning solutions in this area have not been extensively explored. An additional challenge is the scarcity of manually annotated corpora with temporal annotations. To address the adaptability limitations of current systems, we propose a highly adaptable methodology that can be applied to multiple domains and languages. This can be achieved by leveraging a multilingual Pre-trained Language Model (PTLM) with a fill-mask architecture, using a Value Intermediate Representation (VIR) where the temporal expression value format is adjusted to the fill-mask representation. Our approach involves a two-phase training process. Initially, the model is trained with a novel masking policy on a large English biomedical corpus that is automatically annotated with normalized temporal expressions, along with a complementary hand-crafted temporal expressions corpus. This addresses the lack of manually annotated data and helps to achieve sufficient capacity for adaptation to diverse domains or languages. In the second phase, we show how the model can be tailored to different domains and languages employing various techniques, showcasing the versatility of the proposed methodology. This approach significantly outperforms existing systems.

Measuring Grammatical Diversity from Small Corpora: Derivational Entropy Rates, Mean Length of Utterances, and Annotation Invariance

Fermín Moscoso del Prado Martín, Suchir Salhan

In many fields, such as language acquisition, neuropsychology of language, the study of aging, and historical linguistics, corpora are used for estimating the diversity of grammatical structures that are produced during a period by an individual, community, or type of speakers. In these cases, treebanks are taken as representative samples of the syntactic structures that might be encountered. Generalizing the potential syntactic diversity from the structures documented in a small corpus requires careful extrapolation whose accuracy is constrained by the limited size of representative sub-corpora. In this article, I demonstrate --theoretically, and empirically-- that a grammar's derivational entropy and the mean length of the utterances (MLU) it generates are fundamentally linked, giving rise to a new measure, the derivational entropy rate. The mean length of utterances becomes the most practical index of syntactic complexity; I demonstrate that MLU is not a mere proxy, but a fundamental measure of syntactic diversity. In combination with the new derivational entropy rate measure, it provides a theory-free assessment of grammatical complexity. The derivational entropy rate indexes the rate at which different grammatical annotation frameworks determine the grammatical complexity of treebanks. I evaluate the Smoothed Induced Treebank Entropy (SITE) as a tool for estimating these measures accurately, even from very small treebanks. I conclude by discussing important implications of these results for both NLP and human language processing

Session 10: Dialogue and Interactive Systems (Tuesday 29 July, 16:00-17:30, Hall 4/5)

In Prospect and Retrospect: Reflective Memory Management for Long-term Personalized Dialogue Agents

Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, Anand Rajan Iyer, Tianlong Chen, huan liu, Chen-Yu Lee, Tomas Pfister

Large Language Models (LLMs) have made significant progress in open-ended dialogue, yet their inability to retain and retrieve relevant information from long-term interactions limits their effectiveness in applications requiring sustained personalization. External memory mechanisms have been proposed to address this limitation, enabling LLMs to maintain conversational continuity. However, existing approaches struggle with two key challenges. First, rigid memory granularity fails to capture the natural semantic structure of conversations, leading to fragmented and incomplete representations. Second, fixed retrieval mechanisms cannot adapt to diverse dialogue contexts and user interaction patterns. In this work, we propose Reflective Memory Management (RMM), a novel mechanism for long-term dialogue agents, integrating forward- and backward-looking reflections: (1) Prospective Reflection, which dynamically summarizes interactions across granularities—utterances, turns, and sessions—into a personalized memory bank for effective future retrieval, and (2) Retrospective Reflection, which iteratively refines the retrieval in an online reinforcement learning (RL) manner based on LLMs' cited evidence. Experiments show that RMM demonstrates consistent improvement across various metrics and benchmarks. For example, RMM shows more than 10% accuracy improvement over the baseline without memory management on the LongMemEval dataset.

Finding A Voice: Exploring the Potential of African American Dialect and Voice Generation for Chatbots

Sarah E. Finch, Ellie S. Paek, Ikseon Choi, Jinho D. Choi

As chatbots become integral to daily life, personalizing systems is key for fostering trust, engagement, and inclusivity. This study examines how linguistic similarity affects chatbot performance, focusing on integrating African American English (AAE) into virtual agents to better serve the African American community. We develop text-based and spoken chatbots using large language models and text-to-speech technology, then evaluate them with AAE speakers against standard English chatbots. Our results show that while text-based AAE chatbots often underperform, spoken chatbots benefit from an African American voice and AAE elements, improving performance and preference. These findings underscore the complexities of linguistic personalization and the dynamics between text and speech modalities, highlighting technological limitations that affect chatbots' AA speech generation and pointing to promising future research directions.

Session 10: Dialogue and Interactive Systems (Tuesday 29 July, 16:00-17:30, Hall 5X)

First-AID: the first Annotation Interface for grounded Dialogues

Stefano Menini, Daniel Russo, Alessio Palmero Aprosio, Marco Guerini

The swift advancement of Large Language Models (LLMs) has led to their widespread use across various tasks and domains, demonstrating remarkable generalization capabilities. However, achieving optimal performance in specialized tasks often requires fine-tuning LLMs with task-specific resources. The creation of high-quality, human-annotated datasets for this purpose is challenging due to financial constraints and the limited availability of human experts. To address these limitations, we propose First-AID, a novel Human-in-the-Loop (HITL) data collection framework for the knowledge-driven generation of synthetic dialogues using LLM prompting. In particular, our framework implements different strategies of data collection that require different user intervention during dialogue generation to reduce post-editing efforts and enhance the quality of generated dialogues. We also evaluated First-AID on misinformation and hate countering dialogues collection, demonstrating (1) its potential for efficient and high-quality data generation and (2) its adaptability to different practical constraint thanks to the three data collection strategies.

Session 10: Efficient/Low-Resource Methods for NLP (Tuesday 29 July, 16:00-17:30, Hall 4/5)

daDPO: Distribution-Aware DPO for Distilling Conversational Abilities

Zhengze Zhang, Shiqi Wang, Yiqun Shen, Simin Guo, Dahua Lin, Wang Xiaoliang, Nguyen Cam-Tu, Fei Tan

Large language models (LLMs) have demonstrated exceptional performance across various applications, but their conversational abilities decline sharply as model size decreases, presenting a barrier to their deployment in resource-constrained environments. Knowledge distillation (KD) with Direct Preference Optimization (DPO) has emerged as a promising approach to enhance the conversational abilities of smaller models using a larger teacher model. However, current methods primarily focus on ``black-box" KD, which only uses the teacher's responses, overlooking the rich distributional information within the teacher's probability distribution. This paper addresses this gap by introducing daDPO (Distillation-Aware DPO), a novel framework that integrates the teacher's distributional information into DPO distillation while preserving theoretical guarantees. Our framework offers a unified objective that enhances both preference optimization and distribution-based distillation. We provide rigorous theoretical analysis and empirical validation, showing that daDPO outperforms existing methods in restoring performance for pruned models and enhancing smaller models within the same LLM family. Notably, in in-domain evaluation, our method enables a 20% pruned Vicuna1.5-7B to achieve near-teacher performance (-7.3% preference rate), and allows Qwen2.5-1.5B to occasionally outperform its 7b teacher model (14.0% win rate).

PISCO: Pretty Simple Compression for Retrieval-Augmented Generation

Maxime Louis, Hervé Déjean, Stéphane Clinchant

Retrieval-Augmented Generation (RAG) pipelines enhance Large Language Models (LLMs) by retrieving relevant documents, but they face scalability issues due to high inference costs and limited context size. Document compression is a practical solution, but current soft compression methods often suffer from accuracy losses and require extensive pretraining. In this paper, we introduce PISCO, a novel method that achieves a 16x compression rate with minimal accuracy loss (0-3%) across diverse RAG-based question-answering (QA) tasks. Unlike existing approaches, PISCO requires no pretraining or annotated data, relying solely on sequence-level knowledge distillation from document-based questions. With the ability to fine-tune a 7-10B LLM in 24 hours on a single A100 GPU, PISCO offers a highly efficient and scalable solution. We present comprehensive experiments showing that PISCO outperforms existing compression models by 8% in accuracy.

SPILL: Zero-shot Intent Clustering based on Selection and Pooling with Large Language Models

I-Fan Lin, Faegheh Hasibi, Suzan Verberne

In this paper, we propose Selection and Pooling with Large Language Models (SPILL), an intuitive, domain-adaptive method for intent clustering without fine-tuning. Existing embeddings-based clustering methods rely on a few labeled examples or unsupervised fine-tuning to optimize results for each new dataset, which makes them less generalizable to multiple datasets. Our goal is to make these existing embedders more generalizable to new domain datasets without further fine-tuning. Inspired by our theoretical derivation and simulation results on the effectiveness of sampling and pooling techniques, we view the clustering task as a small-scale selection problem. A good solution to this problem is associated with better clustering performance. Accordingly, we propose a two-stage approach: First, for each utterance (referred to as the seed), we derive its embedding using an existing embedder. Then, we apply a distance metric to select a pool of candidates close to the seed. Because the embedder is not optimized for new datasets, in the second stage, we use an LLM to further select utterances from these candidates that share the same intent as the seed. Finally, we pool these selected candidates with the seed to derive a refined embedding for the seed. We found that our method generally outperforms directly using an embedder, and it achieves comparable results to other state-of-the-art studies, even those that use much larger models and require fine-tuning, showing its strength and efficiency. Our results indicate that our method enables existing embedders to be further improved without additional fine-tuning, making them more adaptable to new domain datasets. Additionally, viewing the clustering task as a small-scale selection problem gives the potential of using LLMs to customize clustering tasks according to the user's goals.

Enhanced Data Synthesis for LLM through Reasoning Structures Generated by Hierarchical GFlowNet

Tianpeng Bu, Minying Zhang, Hongtao Duan, Shurui Li, lulu hu, Yu Li

Large language models (LLMs) excel in problem-solving but require training data with diverse reasoning processes. Existing methods mainly optimize instruction-response pairs but lack a systematic design for the underlying reasoning structure. This paper proposes RSS: a Reasoning Structure driven data Synthesis method. We first proactively develop a hierarchical GFlowNet to construct reasoning structures efficiently through a coarse-to-fine directed acyclic graph (DAG) growth process. Then reasoning DAGs are leveraged to actively guide the instruction generation via an iterative suggester-editor workflow and enhance response quality using a structure-aware strategy. Experiments show that LLMs trained on our synthetic datasets achieve 48.50%, 84.00%, 79.90% for AlpacaEval2, GSM8K and HumanEval, outperforming existing data synthesis methods.

Capturing Nuanced Preferences: Preference-Aligned Distillation for Small Language Models

Yanggan Gu, Junzhuo Li, Sirui Huang, Xin Zou, Zhenghua Li, Xuming Hu

Aligning small language models (SLMs) with human values typically involves distilling preference knowledge from large language models (LLMs). However, existing distillation methods model preference knowledge in teacher LLMs by comparing pairwise responses, overlooking the extent of difference between responses. This limitation hinders student SLMs from capturing the nuanced preferences for multiple responses. In this paper, we propose a Preference-Aligned Distillation (PAD) framework, which models teacher's preference knowledge as a probability distribution over all potential preferences, thereby providing more nuanced supervisory signals. Our insight in developing PAD is rooted in the demonstration that language models can serve as reward functions, reflecting their intrinsic preferences. Based on this, PAD comprises three key steps: (1) sampling diverse responses using high-temperature; (2) computing rewards for both teacher and student to construct their intrinsic preference; and (3) training the student's intrinsic preference distribution to align with the teacher's. Experiments on four mainstream alignment benchmarks demonstrate that PAD consistently and significantly outperforms existing approaches, achieving over 20% improvement on AlpacaEval 2 and Arena-Hard, indicating superior alignment with human preferences. Notably, on MT-Bench, using the \textsc{Gemma} model family, the student trained by PAD surpasses its teacher, further validating the effectiveness of our PAD.

Automatic Transmission for LLM Tiers: Optimizing Cost and Accuracy in Large Language Models

Injae Na, Keonwoong Noh, Woohwan Jung

LLM providers typically offer multiple LLM tiers, varying in performance and price. As NLP tasks become more complex and modularized, selecting the suitable LLM tier for each subtask is a key challenge to balance between cost and performance. To address the problem, we introduce LLM Automatic Transmission (LLM-AT) framework that automatically selects LLM tiers without training. LLM-AT consists of Starter, Generator, and Judge. The starter selects the initial LLM tier expected to solve the given question, the generator produces a response using the LLM of the selected tier, and the judge evaluates the validity of the response. If the response is invalid, LLM-AT iteratively upgrades to a higher-tier model, generates a new response, and re-evaluates until a valid response is obtained. Additionally, we propose accuracy estimator, which enables the suitable initial LLM tier selection without training. Given an input question, accuracy estimator estimates the expected accuracy of each LLM tier by computing the valid response rate across top-k similar queries from past inference records. Experiments demonstrate that LLM-AT achieves superior performance while reducing costs, making it a practical solution for real-world applications.

C2KD: Cross-layer and Cross-head Knowledge Distillation for Small Language Model-based Recommendation

Xiao Chen, Changyi Ma, Wenqi Fan, Zhaoxiang Zhang, Li Qing

Sequential recommenders predict users' next interactions based on historical behavior and are essential in modern recommendation systems. While Large Language Models (LLMs) show promise, their size and high inference costs limit deployment on resource-constrained devices. Small Language Models (SLMs) provide a more efficient alternative for edge devices, but bridging the recommendation performance gap between LLMs and SLMs remains challenging. Typical approaches like supervised fine-tuning or vanilla knowledge distillation (KD) often lead to suboptimal performance or even negative transfer. Our motivational experiments reveal key issues with vanilla KD methods: feature imitation suffers from redundancy and uneven recommendation ability across layers, while prediction mimicking faces conflicts caused by differing weight distributions of prediction heads. To address these challenges, we propose a simple yet effective framework, \text{C}^2KD, to transfer task-relevant knowledge from two complementary dimensions. Specifically, our method incorporates: (1) cross-layer feature imitation, which uses a dynamic router to select the most relevant teacher layers and assimilate task-relevant knowledge from the teacher’s late layers, allowing the student to concentrate on the teacher’s specialized knowledge; and (2) cross-head logit distillation, which maps the intermediate features of the student to the teacher’s output head, thereby minimizing prediction discrepancies between the teacher and the student. Extensive experiments across diverse model families demonstrate that our approach enables 1B-parameter SLMs to achieve competitive performance compared to LLMs (e.g., Llama3-8B), offering a practical solution for real-world on-device sequential recommendations.

On the Generalization vs Fidelity Paradox in Knowledge Distillation

Suhas Kamasetty Ramesh, Ayan Sengupta, Tanmoy Chakraborty

Knowledge distillation (KD) is a key technique for compressing large language models into smaller ones while preserving performance. Despite the recent traction of KD research, its effectiveness for smaller language models (LMs) and the mechanisms driving knowledge transfer remain underexplored. In this work, we present the first large-scale empirical and statistical analysis of KD across models ranging from 0.5B to 7B parameters on 14 complex reasoning tasks in a zero-shot setting. Our findings reveal that KD can improve the average performance of smaller models by up to 10%, with a peak task specific gain of 22%, while providing only marginal benefits (\sim 1.3%) for larger models. Surprisingly, teacher performance has a minimal impact on student outcomes, while teacher task expertise impacts KD effectiveness. A correlation study indicates that smaller LMs benefit more from KD, whereas larger LMs show diminished gains. Additionally, we uncover a misalignment between improvements in student performance and reasoning fidelity, suggesting that while KD enhances accuracy, it does not always maintain the structured decision-making processes of the teacher. Our ablation study further highlights the importance of teacher signals and logit smoothing in influencing students' performance after distillation. Overall, our study offers a comprehensive empirical and statistical assessment of KD, highlighting both its benefits and trade-offs when distilling knowledge from larger to smaller LMs.

FPE2M2: Approaching Lossless and Efficient Quantization with Native Floating Point

Ke Yi, jianwei zhang, Zhiying Xu, Xinlong Yang, Yang Zhou, Minmin Sun, Zengke Liu, Tong Zhang, Junyang Lin, Jingren Zhou

Auto-regressive decoding is a memory-bound job, meaning decoding inference performance is limited by the bandwidth rather than the computational capabilities of the GPU. Weight-only quantization is a promising method to address the memory-bound limitations. Previous studies have followed one of two approaches. Some have exclusively studied integer quantization while ignoring the Gaussian distribution nature of LLMs' weights. Others have proposed non-uniform quantization but incurred additional I/O overhead due to lookup tables, e.g. NF4. In this work, we extend the IEEE 754 float-point standard to the ExMy quantization schema, which allocates x bit for the exponent and y bit for the mantissa to represent a number. In terms of runtime efficiency, we demonstrate that the conversion from ExMy to FP16 can be realized through register-level operations, which can get almost the same performance as INT5. In terms of quantization loss, we analyze that of different ExMy settings, where the E2M2 schema achieves an optimal balance, offering the highest efficiency with lossless accuracy. We further propose the FPE2M2 framework that supports lossless weight-only quantization inference and validate the FPE2M2 framework on Qwen and LLaMA Models across various modalities, such as text, image, and audio tasks, which achieves a faster inference speed while maintaining nearly lossless accuracy.

MiniKV: Pushing the Limits of 2-Bit KV Cache via Compression and System Co-Design for Efficient Long Context Inference

Akshat Sharma, Hangliang Ding, Jianping Li, Neel Dani, Minjia Zhang

State-of-the-art 2-bit KV cache quantization techniques achieve excellent results in accelerating LLM inference while retaining accuracy on long context tasks. However, further pushing the compression ratio fails to deliver performance gains. In this work, we revisit these approaches by considering, additionally, adaptive KV methods that retain LLM accuracy with only a subset of KV states. This leads us to propose a method based on 2-bit KV cache quantization with adaptive KV policies. In addition, we take an algorithm and system co-design approach by developing hardware-friendly kernels to accelerate LLM inference while making MiniKV compatible with existing memory-efficient attention techniques such as FlashAttention, effectively translating algorithmic improvements into system performance gains. Experiments on a wide range of long context tasks show that MiniKV effectively achieves >80% KV cache compression while retaining accuracy, outperforming state-of-the-art methods while achieving excellent latency, throughput, and memory consumption improvements in long context inference.

Adaptive LoRA Merge with Parameter Pruning for Low-Resource Generation

Ryota Miyano, Yuki Arase

This study proposes a simple yet effective LoRA merge method to achieve LLM adaptation for low-resource language generation tasks. The LoRA merge technique, which integrates multiple LoRA modules trained on different tasks, has gained attention as an effective and efficient approach for adapting LLMs to target tasks. However, previous methods are limited in adaptability as they keep the LoRA parameters frozen. Additionally, the low-resource problem has been out of their scope. We propose a LoRA merge method that updates LoRA parameters during the merge process to improve task-specific adaptability. Furthermore, we further enhance the adaptability by parameter pruning that allows finer-grained adjustments of LoRA parameters during merging. Extensive experiments have been conducted taking summarization as a benchmark task. Our datasets cover various domains and multiple languages of English and Japanese. The results confirm that the proposed method achieves significant and consistent improvements in task adaptability over the previous methods.

Direct Behavior Optimization: Unlocking the Potential of Lightweight LLMs

Hongming Yang, Shi Lin, Jun Shao, Changting Lin, Donghai Zhu, Meng Han, Qinglei Kong

Lightweight Large Language Models (LwLLMs) are reduced-parameter, optimized models designed to run efficiently on consumer-grade hardware, offering significant advantages in resource efficiency, cost-effectiveness, and data privacy. However, these models often struggle with limited inference and reasoning capabilities, which restrict their performance on complex tasks and limit their practical applicability. Moreover, existing prompt optimization methods typically rely on extensive manual effort or the meta-cognitive abilities of state-of-the-art LLMs, making them less effective for LwLLMs. To address these challenges, we introduce DeBoP, a new Direct Behavior Optimization Paradigm, original from the Chain-of-Thought (CoT) prompting technique. Unlike CoT Prompting, DeBoP is an automatic optimization method, which focuses on the optimization directly on the behavior of LwLLMs. In particular, DeBoP transforms the optimization of complex prompts into the optimization of discrete, quantifiable execution sequences using a gradient-free Monte Carlo Tree Search. We evaluate DeBoP on seven challenging tasks where state-of-the-art LLMs excel but LwLLMs generally underperform. Experimental results demonstrate that DeBoP significantly outperforms recent prompt optimization methods on most tasks. In particular, DeBoP-optimized LwLLMs surpass GPT-3.5 on most tasks while reducing computational time by approximately 60% compared to other automatic prompt optimization methods.

DReSD: Dense Retrieval for Speculative Decoding

Milan Gritta, Huiyin Xue, Gerasimos Lampouras

Speculative decoding (SD) accelerates Large Language Model (LLM) generation by using an efficient draft model to propose the next few tokens, which are verified by the LLM in a single forward call, reducing latency while preserving its outputs. We focus on retrieval-based SD where the draft model retrieves the next tokens from a non-parametric datastore. Sparse retrieval \cite[REST], which operates on the surface form of strings, is currently the dominant paradigm due to its simplicity and scalability. However, its effectiveness is limited due to the usage of short contexts and exact string matching. Instead, we introduce Dense Retrieval for Speculative Decoding (DReSD), a novel framework that uses approximate nearest neighbour search with contextualised token embeddings to retrieve the most semantically relevant token sequences for SD. Extensive experiments show that DReSD achieves (on average) 87% higher acceptance rates, 65% longer accepted tokens and 19% faster generation speeds compared to sparse retrieval (REST).

TagRouter: Learning Route to LLMs through Tags for Open-Domain Text Generation Tasks

Zhou Chen, Zhiqiang wei, Yuqi Bai, Xue Xiong, Jianmin WU

Model routing allocates queries to the suitable model, improving system performance while reducing costs. However, existing routing methods face practical limitations that hinder scalability in large-scale applications and struggle to keep up with the rapid growth of the large language model (LLM) ecosystem. To tackle these challenges, we propose TagRouter, a training-free model routing method designed to optimize the synergy among multiple LLMs for open-domain text generation tasks. Experimental results demonstrate that TagRouter outperforms 13 baseline methods, increasing the accept rate of system by 6.15% and reducing costs by 17.20%, achieving optimal cost-efficiency. Our findings provides the LLM community with an efficient and scalable solution for model ensembling, offering users an evolvable "super model."

Small Encoders Can Rival Large Decoders in Detecting Groundedness

Istabrak Abbes, Gabriele Prato, Quentin Fournier, Fernando Rodriguez, Alaa Boukhary, Adam Elwood, Sarath Chandar

Augmenting large language models (LLMs) with external context significantly improves their performance in natural language processing (NLP) tasks. However, LLMs struggle to answer queries reliably when the provided context lacks information, often resorting to ungrounded speculation or internal knowledge. Groundedness – generating responses strictly supported by the context – is essential for ensuring factual consistency and trustworthiness. This study focuses on detecting whether a given query is grounded in a document provided in context before the costly answer generation by LLMs. Such a detection mechanism can significantly reduce both inference time and resource consumption. We show that lightweight, task-specific encoder models such as RoBERTa and NomicBERT, fine-tuned on curated datasets, can achieve accuracy comparable to state-of-the-art LLMs, such as Llama3 8B and GPT4o, in groundedness detection while reducing inference latency by orders of magnitude.

FREE: Fast and Robust Vision Language Models with Early Exits

Divya Jyoti Bajpai, Manjesh Kumar Hanawal

In recent years, Vision-Language Models (VLMs) have shown remarkable performance improvements in Vision-Language tasks. However, their large size poses challenges for real-world applications where inference latency is a concern. To tackle this issue, we propose employing Early Exit (EE) strategies in VLMs. However, training exit classifiers in VLMs is challenging, particularly with limited labeled training data. To address this, we introduce FREE, an adversarial training approach within a GAN-based framework. Here, each exit consists of a transformer layer and a classifier. The transformer layer is adversarially trained to produce feature representations similar to the final layer, while a feature classifier serves as the discriminator. Our method focuses on performing input-adaptive inference that increases inference speed with minimal drop in performance. Experimental results demonstrate the effectiveness of our approach in enhancing accuracy and model robustness by mitigating overthinking and the phenomenon of mid-crisis that we highlight. We experimentally validate that our method speeds up the inference process by more than 1.51\times while retaining comparable performance. The anonymized source code is available at https://github.com/Div290/BLIPEE.

Small Models Struggle to Learn from Strong Reasoners

Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, Radha Poovendran

Large language models (LLMs) excel in complex reasoning tasks, and distilling their reasoning capabilities into smaller models has shown promise. However, we uncover an interesting phenomenon, which we term the Small Model Learnability Gap: small models (3B parameters) do not consistently benefit from long chain-of-thought (CoT) reasoning or distillation from larger models. Instead, they perform better when fine-tuned on shorter, simpler reasoning chains that better align with their intrinsic learning capacity. To address this, we propose Mix Distillation, a simple yet effective strategy that balances reasoning complexity by combining long and short CoT examples or reasoning from both larger and smaller models. Our experiments demonstrate that Mix Distillation significantly improves small model reasoning performance compared to training on either data alone. These findings highlight the limitations of direct strong model distillation and underscore the importance of adapting reasoning complexity for effective reasoning capability transfer.

Dictionaries to the Rescue: Cross-Lingual Vocabulary Transfer for Low-Resource Languages Using Bilingual Dictionaries

Haruki Sakajo, Yusuke Ide, Justin Vasselli, Yusuke Sakai, Yingtao Tian, Hidetaka Kamigaito, Taro Watanabe

Cross-lingual vocabulary transfer plays a promising role in adapting pre-trained language models to new languages, including low-resource languages. Existing approaches that utilize monolingual or parallel corpora face challenges when applied to languages with limited resources. In this work, we propose a simple yet effective vocabulary transfer method that utilizes bilingual dictionaries, which are available for many languages, thanks to descriptive linguists. Our proposed method leverages a property of BPE tokenizers where removing a subword from the vocabulary causes a fallback to shorter subwords. The embeddings of target subwords are estimated iteratively by progressively removing them from the tokenizer. The experimental results show that our approach outperforms existing methods for low-resource languages, demonstrating the effectiveness of a dictionary-based approach for cross-lingual vocabulary transfer.

L4Q: Parameter Efficient Quantization-Aware Fine-Tuning on Large Language Models

Hyesung Jeon, Yulhwa Kim, Jae-Joon Kim

Due to the high memory and computational costs associated with large language models (LLMs), model compression techniques such as quantization, which reduces inference costs, and parameter-efficient fine-tuning (PEFT) methods like Low-Rank Adaptation (LoRA), which reduce training costs, have gained significant popularity. This trend has spurred active research into quantization-aware PEFT techniques, aimed at maintaining model accuracy while minimizing memory overhead during both inference and training. Previous quantization-aware PEFT methods typically apply post-training quantiation (PTQ) to pre-trained LLMs, followed by PEFT to recover accuracy loss. Meanwhile, this approach has limitations in recovering the accuracy loss. In this paper, we propose L4Q, a method that integrates Quantization-Aware Training (QAT) with LoRA. By employing a memory-optimized layer design, L4Q significantly reduces QAT's memory overhead, making its training cost comparable to LoRA, while preserving the advantage of QAT in producing fully quantized LLMs with high accuracy. Our experiments demonstrate that this combined approach to quantization and fine-tuning achieves superior accuracy compared to decoupled fine-tuning schemes, particularly in 4-bit and 3-bit quantization, positioning L4Q as an efficient QAT solution. Using the LLaMA and Mistral models with instructional datasets, we showcase L4Q's capabilities in language tasks and few-shot learning.

Prompt Candidates, then Distill: A Teacher-Student Framework for LLM-driven Data Annotation

Mingxuan Xia, Haobo Wang, Yixuan Li, Zewei Yu, Jindong Wang, Junbo Zhao, Runze Wu

Recently, Large Language Models (LLMs) have demonstrated significant potential for data annotation, markedly reducing the labor costs associated with downstream applications. However, existing methods mostly adopt an aggressive strategy by prompting LLM to determine a single gold label for each unlabeled sample. Due to the inherent uncertainty within LLMs, they often produce incorrect labels for difficult samples, severely compromising the data quality for downstream applications. Motivated by ambiguity aversion in human behaviors, we propose a novel candidate annotation paradigm wherein large language models are encouraged to output all possible labels when incurring uncertainty. To ensure unique labels are provided for downstream tasks, we develop a teacher-student framework CanDist that distills candidate annotations with a Small Language Model (SLM). We further provide a rigorous justification demonstrating that distilling candidate annotations from the teacher LLM offers superior theoretical guarantees compared to directly using single annotations. Extensive experiments across six text classification tasks validate the effectiveness of our proposed method. The source code is available at https://github.com/MingxuanXia/CanDist.

Enhancing Unsupervised Sentence Embeddings via Knowledge-Driven Data Augmentation and Gaussian-Decayed Contrastive Learning

Peichao Lai, Zhengfeng Zhang, Wentao Zhang, Fangcheng Fu, Bin CUI

Recently, using large language models (LLMs) for data augmentation has led to considerable improvements in unsupervised sentence embedding models. However, existing methods encounter two primary challenges: limited data diversity and high data noise. Current approaches often neglect fine-grained knowledge, such as entities and quantities, leading to insufficient diversity. Besides, unsupervised data frequently lacks discriminative information, and the generated synthetic samples may introduce noise. In this paper, we propose a pipeline-based data augmentation method via LLMs and introduce the Gaussian-decayed gradient-assisted Contrastive Sentence Embedding (GCSE) model to enhance unsupervised sentence embeddings. To tackle the issue of low data diversity, our pipeline utilizes knowledge graphs (KGs) to extract entities and quantities, enabling LLMs to generate more diverse samples. To address high data noise, the GCSE model uses a Gaussian-decayed function to limit the impact of false hard negative samples, enhancing the model's discriminative capability. Experimental results show that our approach achieves state-of-the-art performance in semantic textual similarity (STS) tasks, using fewer data samples and smaller LLMs, demonstrating its efficiency and robustness across various models.

Plug-in and Fine-tuning: Bridging the Gap between Small Language Models and Large Language Models

Kyeonghyun Kim, Jinhee Jang, Juhwan Choi, Yoonji Lee, Kyohoon Jin, YoungBin Kim

Large language models (LLMs) are renowned for their extensive linguistic knowledge and strong generalization capabilities, but their high computational demands make them unsuitable for resource-constrained environments. In contrast, small language models (SLMs) are computationally efficient but often lack the broad generalization capacity of LLMs. To bridge this gap, we propose PiFi, a novel framework that combines the strengths of both LLMs and SLMs to achieve high performance while maintaining efficiency. PiFi integrates a single frozen layer from an LLM into a SLM and fine-tunes the combined model for specific tasks, boosting performance without a significant increase in computational cost. We show that PiFi delivers consistent performance improvements across a range of natural language processing tasks, including both natural language understanding and generation. Moreover, our findings demonstrate PiFi's ability to effectively leverage LLM knowledge, enhancing generalization to unseen domains and facilitating the transfer of linguistic abilities.

CORAL: Learning Consistent Representations across Multi-step Training with Lighter Speculative Drafter

Yepeng Weng, Dianwen Mei, Huishi Qiu, Xujie Chen, Li Liu, Jiang Tian, Zhongchao Shi

Speculative decoding is a powerful technique that accelerates Large Language Model (LLM) inference by leveraging a lightweight speculative draft model. However, existing designs suffers in performance due to misalignment between training and inference. Recent methods have tried to solve this issue by adopting a multi-step training strategy, but the complex inputs of different training steps make it harder for the draft model to converge. To address this, we propose CORAL, a novel framework that improves both accuracy and efficiency in speculative drafting. CORAL introduces Cross-Step Representation Alignment, a method that enhances consistency across multiple training steps, significantly improving speculative drafting performance. Additionally, we identify the LM head as a major bottleneck in the inference speed of the draft model. We introduce a weight-grouping mechanism that selectively activates a subset of LM head parameters during inference, substantially reducing the latency of the draft model. We evaluate CORAL on three LLM families and three benchmark datasets, achieving speedup ratios of 2.50x-4.07x, outperforming state-of-the-art methods such as EAGLE-2 and HASS. Our results demonstrate that CORAL effectively mitigates training-inference misalignment and delivers significant speedup for modern LLMs with large vocabularies.

Limited-Resource Adapters Are Regularizers, Not Linguists

Marcell Fekete, Nathaniel Romney Robinson, Ernests Lavrinovics, Djeride Jean-Baptise, Raj Dabre, Johannes Bjerva, Heather Lent

Cross-lingual transfer from related high-resource languages is a well-established strategy to enhance low-resource language technologies. Prior work has shown that adapters show promise for, e.g., improving low-resource machine translation (MT). In this work, we investigate an adapter souping method combined with cross-attention fine-tuning of a pre-trained MT model to leverage language transfer for three low-resource Creole languages, which exhibit relatedness to different language groups across distinct linguistic dimensions. Our approach improves performance substantially over baselines. However, we find that linguistic relatedness---or even a lack thereof---does not covary meaningfully with adapter performance. Surprisingly, our cross-attention fine-tuning approach appears equally effective with randomly initialized adapters, implying that the benefit of adapters in this setting lies in parameter regularization, and not in meaningful information transfer. We provide analysis supporting this regularization hypothesis. Our findings underscore the reality that neural language processing involves many success factors, and that not all neural methods leverage linguistic knowledge in intuitive ways.

GIFT-SW: Gaussian noise Injected Fine-Tuning of Salient Weights for LLMs

Maxim Zhelnin, Viktor Moskvoretskii, Egor Shvetsov, Maria Krylova, Venediktov Egor, Zuev Aleksandr, Evgeny Burnaev

Parameter Efficient Fine-Tuning (PEFT) methods have gained popularity and democratized the usage of Large Language Models (LLMs). Recent studies have shown that a small subset of weights significantly impacts performance. Based on this observation, we introduce a novel PEFT method, called Gaussian noise Injected Fine Tuning of Salient Weights (GIFT-SW). Our method updates only salient columns, while injecting Gaussian noise into non-salient ones. To identify these columns, we developed a generalized sensitivity metric that extends and unifies metrics from previous studies. Experiments with LLaMA models demonstrate that GIFT-SW outperforms full fine-tuning and modern PEFT methods under the same computational budget. Moreover, GIFT-SW offers practical advantages to recover performance of models subjected to mixed-precision quantization with keeping salient weights in full precision.

Quaff: Quantized Parameter-Efficient Fine-Tuning under Outlier Spatial Stability Hypothesis

Hong Huang, Dapeng Wu

Large language models (LLMs) have made exciting achievements across various domains, yet their deployment on resource-constrained personal devices remains hindered by the prohibitive computational and memory demands of task-specific fine-tuning. While quantization offers a pathway to efficiency, existing methods struggle to balance performance and overhead, either incurring high computational/memory costs or failing to address activation outliers—a critical bottleneck in quantized fine-tuning. To address these challenges, we propose the Outlier Spatial Stability Hypothesis (__OSSH__): _During fine-tuning, certain activation outlier channels retain stable spatial positions across training iterations._ Building on OSSH, we propose __Quaff__, a Quantized parameter-efficient fine-tuning framework for LLMs, optimizing low-precision activation representations through targeted momentum scaling. Quaff dynamically suppresses outliers exclusively in invariant channels using lightweight operations, eliminating full-precision weight storage and global rescaling while reducing quantization errors. Extensive experiments across ten benchmarks validate OSSH and demonstrate Quaff's efficacy. Specifically, on the GPQA reasoning benchmark, Quaff achieves a 1.73\times latency reduction and 30% memory savings over full-precision fine-tuning while improving accuracy by 0.6% on the Phi-3 model, reconciling the triple trade-off between efficiency, performance, and deployability. By enabling consumer-grade GPU fine-tuning (e.g., RTX 2080 Super) without sacrificing model utility, Quaff democratizes personalized LLM deployment. The code is available at https://anonymous.4open.science/r/Quaff-B322/.

Efficient Pretraining Data Selection for Language Models via Multi-Actor Collaboration

Tianyi Bai, Ling Yang, Zhen Hao Wong, Fupeng Sun, Xinlin Zhuang, Jiahui Peng, Chi Zhang, Lijun Wu, Qiu Jiantao, Wentao Zhang, Binhang Yuan, Conghui He

Efficient data selection is crucial to accelerate the pretraining of language model (LMs). While various methods have been proposed to enhance data efficiency, limited research has addressed the inherent conflicts between these approaches to achieve optimal data selection for LM pretraining. To tackle this problem, we propose a multi-actor collaborative data selection mechanism. Each data selection method independently prioritizes data based on its specific criterion and updates its prioritization rules using the current state of the model, functioning as an independent actor for data selection. Additionally, a console is designed to adjust the impacts of different actors at various stages and dynamically integrate information from all actors throughout the LM pretraining process. We conduct extensive empirical studies to evaluate our multi-actor framework. The experimental results demonstrate that our approach significantly improves data efficiency, accelerates convergence in LM pretraining, and achieves an average relative performance gain up to 10.5% across multiple language model benchmarks compared to the state-of-the-art methods.

DenseLoRA: Dense Low-Rank Adaptation of Large Language Models

Lin Mu, Xiaoyu Wang, Li Ni, Yang Li, Zhize Wu, Peiquan Jin, Yiwen Zhang

Low-rank adaptation (LoRA) has been developed as an efficient approach for adapting large language models (LLMs) by fine-tuning two low-rank matrices, thereby reducing the number of trainable parameters. However, prior research indicates that many of the weights in these matrices are redundant, leading to inefficiencies in parameter utilization. To address this limitation, we introduce Dense Low-Rank Adaptation (DenseLoRA), a novel approach that enhances parameter efficiency while achieving superior performance compared to LoRA. DenseLoRA builds upon the concept of representation fine-tuning, incorporating a single Encoder-Decoder to refine and compress hidden representations across all adaptation layers before applying adaptation. Instead of relying on two redundant low-rank matrices as in LoRA, DenseLoRA adapts LLMs through a dense low-rank matrix, improving parameter utilization and adaptation efficiency. We evaluate DenseLoRA on various benchmarks, showing that it achieves 83.8% accuracy with only 0.01% of trainable parameters, compared to LoRA's 80.8% accuracy with 0.70% of trainable parameters on LLaMA3-8B. Additionally, we conduct extensive experiments to systematically assess the impact of DenseLoRA's components on overall model performance.

MoQAE: Mixed-Precision Quantization for Long-Context LLM Inference via Mixture of Quantization-Aware Experts

Wei Tao, Haocheng Lu, Xiaoyang Qu, Bin Zhang, Kai Lu, Jiguang Wan, Jianzong Wang

One of the primary challenges in optimizing large language models (LLMs) for long-context inference lies in the high memory consumption of the Key-Value (KV) cache. Existing approaches, such as quantization, have demonstrated promising results in reducing memory usage. However, current quantization methods cannot take both effectiveness and efficiency into account. In this paper, we propose MoQAE, a novel mixed-precision quantization method via mixture of quantization-aware experts. First, we view different quantization bit-width configurations as experts and use the traditional mixture of experts (MoE) method to select the optimal configuration. To avoid the inefficiency caused by inputting tokens one by one into the router in the traditional MoE method, we input the tokens into the router chunk by chunk. Second, we design a lightweight router-only fine-tuning process to train MoQAE with a comprehensive loss to learn the trade-off between model accuracy and memory usage. Finally, we introduce a routing freezing (RF) and a routing sharing (RS) mechanism to further reduce the inference overhead. Extensive experiments on multiple benchmark datasets demonstrate that our method outperforms state-of-the-art KV cache quantization approaches in both efficiency and effectiveness.

UORA: Uniform Orthogonal Reinitialization Adaptation in Parameter Efficient Fine-Tuning of Large Models

Xueyan Zhang, Jinman Zhao, Zhifei Yang, Yibo Zhong, Shuhao Guan, Linbo Cao, Yining Wang

This paper introduces UoRA, a novel parameter-efficient fine-tuning (PEFT) approach for large language models (LLMs). UoRA achieves state-of-the-art efficiency by leveraging a low-rank approximation method that reduces the number of trainable parameters without compromising performance. Unlike existing methods such as LoRA and VeRA, UoRA employs a re-parametrization mechanism that eliminates the need to adapt frozen projection matrices while maintaining shared projection layers across the model. This results in halving the trainable parameters compared to LoRA and outperforming VeRA in computation and storage efficiency. Comprehensive experiments across various benchmarks demonstrate UoRA's superiority in achieving competitive fine-tuning performance with minimal computational overhead. We demonstrate its performance on GLUE and E2E benchmarks and is effectiveness in instruction-tuning large language models and image classification models. Our contributions establish a new paradigm for scalable and resource-efficient fine-tuning of LLMs.

BeamLoRA: Beam-Constraint Low-Rank Adaptation

Naibin Gu, Zhenyu Zhang, Xiyu Liu, Peng Fu, Zheng Lin, Shuohuan Wang, Yu Sun, Hua Wu, Weiping Wang, Haifeng Wang

Due to the demand for efficient fine-tuning of large language models, Low-Rank Adaptation (LoRA) has been widely adopted as one of the most effective parameter-efficient fine-tuning methods. Nevertheless, while LoRA improves efficiency, there remains room for improvement in accuracy. Herein, we adopt a novel perspective to assess the characteristics of LoRA ranks. The results reveal that different ranks within the LoRA modules not only exhibit varying levels of importance but also evolve dynamically throughout the fine-tuning process, which may limit the performance of LoRA. Based on these findings, we propose BeamLoRA, which conceptualizes each LoRA module as a beam where each rank naturally corresponds to a potential sub-solution, and the fine-tuning process becomes a search for the optimal sub-solution combination. BeamLoRA dynamically eliminates underperforming sub-solutions while expanding the parameter space for promising ones, enhancing performance with a fixed rank. Extensive experiments across three base models and 12 datasets spanning math reasoning, code generation, and commonsense reasoning demonstrate that BeamLoRA consistently enhances the performance of LoRA, surpassing the other baseline methods.

Phonotomizer: A Compact, Unsupervised, Online Training Approach to Real-Time, Multilingual Phonetic Segmentation

Michael S. Yantosca, Albert M. K. Cheng

Phonetic transcription requires significant time and expert training. Automated, state-of-the-art text-dependent methods still involve substantial pre-training annotation labor and may not generalize to multiple languages. Hallucination of speech amid silence or non-speech noise can also plague these methods, which fall short in real-time applications due to post hoc whole-phrase evaluation. This paper introduces Phonotomizer, a compact, unsupervised, online training approach to automatic, multilingual phonetic segmentation, a critical first stage in transcription. Unlike prior approaches, Phonotomizer trains on raw sound files alone and can modulate computational exactness. Preliminary evaluations on Irish and Twi, two underrepresented languages, exhibit segmentation comparable to current forced alignment technology, reducing acoustic model size and minimizing training epochs.

EAC-MoE: Expert-Selection Aware Compressor for Mixture-of-Experts Large Language Models

Yuanteng Chen, Yuantian Shao, Peisong Wang, Jian Cheng

Mixture-of-Experts (MoE) has demonstrated promising potential in scaling LLMs. However, it is hindered by two critical challenges: (1) substantial GPU memory consumption to load all experts; (2) low activated parameters cannot be equivalently translated into inference acceleration effects. In this work, we propose EAC-MoE, an Expert-Selection Aware Compressor for MoE-LLMs, which deeply aligns with the characteristics of MoE from the perspectives of quantization and pruning, and introduces two modules to address these two challenges respectively: (1) The expert selection bias caused by low-bit quantization is a major factor contributing to the performance degradation in MoE-LLMs. Based on this, we propose Quantization with Expert-Selection Calibration (QESC), which mitigates the expert selection bias by calibrating the routers within the MoE; (2) There are always certain experts that are not crucial for the corresponding tasks, yet causing inference latency. Therefore, we propose Pruning based on Expert-Selection Frequency (PESF), which significantly improves inference speed by pruning less frequently used experts for current task. Extensive experiments demonstrate that our approach significantly reduces memory usage and improves inference speed with minimal performance degradation.

From Selection to Generation: A Survey of LLM-based Active Learning

Yu Xia, Subhojyoti Mukherjee, Zhouhang Xie, Junda Wu, Xintong Li, Ryan Aponte, Hanjia Lyu, Joe Barrow, Hongjie Chen, Franck Dernoncourt, Branislav Kveton, Tong Yu, Ruiyi Zhang, Jiuxiang Gu, Nesreen K. Ahmed, Yu Wang, Xiang Chen, Hanieh Deilamsalehy, Sungchul Kim, Zhengmian Hu, Yue Zhao, Nedim Lipka, Seunghyun Yoon, Ting-Hao Kenneth Huang, Zichao Wang, Puneet Mathur, Soumyabrata Pal, Koyel Mukherjee, Zhehao Zhang, Namyong Park, Thien Huu Nguyen, Jiebo Luo, Ryan A. Rossi, Julian McAuley

Active Learning (AL) has been a powerful paradigm for improving model efficiency and performance by selecting the most informative data points for labeling and training. In recent active learning frameworks, Large Language Models (LLMs) have been employed not only for selection but also for generating entirely new data instances and providing more cost-effective annotations. Motivated by the increasing importance of high-quality data and efficient model training in the era of LLMs, we present a comprehensive survey on LLM-based Active Learning. We introduce an intuitive taxonomy that categorizes these techniques and discuss the transformative roles LLMs can play in the active learning loop. We further examine the impact of AL on LLM learning paradigms and its applications across various domains. Finally, we identify open challenges and propose future research directions. This survey aims to serve as an up-to-date resource for researchers and practitioners seeking to gain an intuitive understanding of LLM-based AL techniques and deploy them to new applications.

FEAT: A Preference Feedback Dataset through a Cost-Effective Auto-Generation and Labeling Framework for English AI Tutoring

Hyein Seo, Taewook Hwang, Yohan Lee, Sangkeun Jung

In English education tutoring, teacher feedback is essential for guiding students. Recently, AI-based tutoring systems have emerged to assist teachers; however, these systems require high-quality and large-scale teacher feedback data, which is both time-consuming and costly to generate manually. In this study, we propose FEAT, a cost-effective framework for generating teacher feedback, and have constructed three complementary datasets: (1) DIRECT-Manual (DM), where both humans and large language models (LLMs) collaboratively generate high-quality teacher feedback, albeit at a higher cost; (2) DIRECT-Generated (DG), an LLM-only generated, cost-effective dataset with lower quality;, and (3) DIRECT-Augmented (DA), primarily based on DG with a small portion of DM added to enhance quality while maintaining cost-efficiency. Experimental results showed that incorporating a small portion of DM (5–10%) into DG leads to superior performance compared to using 100% DM alone.

Improving Chain-of-Thought Reasoning via Quasi-Symbolic Abstractions

Leonardo Ranaldi, Marco Valentino, Andre Freitas

Chain-of-Though (CoT) represents a common strategy for reasoning in Large Language Models (LLMs) by decomposing complex tasks into intermediate inference steps. However, explanations generated via CoT are susceptible to content biases that negatively affect their robustness and faithfulness. To mitigate existing limitations, recent work has proposed using logical formalisms coupled with external symbolic solvers. However, fully symbolic approaches possess the bottleneck of requiring a complete translation from natural language to formal languages, a process that affects efficiency and flexibility. To achieve a trade-off, this paper investigates methods to disentangle content from logical reasoning without a complete formalisation. In particular, we present QuaSAR (for Quasi-Symbolic Abstract Reasoning), a variation of CoT that guides LLMs to operate at a higher level of abstraction via quasi-symbolic explanations. Our framework leverages the capability of LLMs to formalise only relevant variables and predicates, enabling the coexistence of symbolic elements with natural language. We show the impact of QuaSAR for in-context learning and for constructing demonstrations to improve the reasoning capabilities of smaller models. Our experiments show that quasi-symbolic abstractions can improve CoT-based methods by up to 8% accuracy, enhancing robustness and consistency on challenging adversarial variations on both natural language (i.e. MMLU-Redux) and symbolic reasoning tasks (i.e., GSM-Symbolic).

Multilingual Encoder Knows more than You Realize: Shared Weights Pretraining for Extremely Low-Resource Languages

Zeli Su, Ziyin Zhang, Guixian Xu, Jianing Liu, Xu Han, Ting Zhang, Yushuang Dong

While multilingual language models like XLM-R have advanced multilingualism in NLP, they still perform poorly in extremely low-resource languages. This situation is exacerbated by the fact that modern LLMs such as LLaMA and Qwen support far fewer languages than XLM-R, making text generation models non-existent for many languages in the world. To tackle this challenge, we propose a novel framework for adapting multilingual encoders to text generation in extremely low-resource languages. By reusing the weights between the encoder and the decoder, our framework allows the model to leverage the learned semantic space of the encoder, enabling efficient learning and effective generalization in low-resource languages. Applying this framework to four Chinese minority languages, we present XLM-SWCM, and demonstrate its superior performance on various downstream tasks even when compared with much larger models.

Controlled Low-Rank Adaptation with Subspace Regularization for Continued Training on Large Language Models

Yuheng Lu, Bingshuo Qian, Caixia Yuan, Huixing Jiang, Xiaojie Wang

Large language models (LLMs) exhibit remarkable capabilities in natural language processing but face catastrophic forgetting when learning new tasks, where adaptation to a new domain leads to a substantial decline in performance on previous tasks. In this paper, we propose Controlled LoRA (CLoRA), a subspace regularization method on LoRA structure. Aiming to reduce the scale of output change while introducing minimal constraint on model capacity, CLoRA imposes constraints on the direction of updating matrix's null space. Experimental results on one-stage LLM finetuning tasks and continual learning settings highlight the superiority of CLoRA as an effective parameter-efficient finetuning method with catastrophic forgetting mitigating. Further investigation for model parameters indicates that CLoRA effectively balances the trade-off between model capacity and degree of forgetting. The code for implementing CLoRA will be publicly available.

DAC: A Dynamic Attention-aware Approach for Task-Agnostic Prompt Compression

Yi Zhao, Zuchao Li, hai zhao, Baoyuan Qi, Liu Guoming

Task-agnostic prompt compression leverages the redundancy in natural language to reduce computational overhead and enhance information density within prompts, especially in long-context scenarios. Existing methods predominantly rely on information entropy as the metric to compress lexical units, aiming to achieve minimal information loss. However, these approaches overlook two critical aspects: (i) the importance of attention-critical tokens at the algorithmic level, and (ii) shifts in information entropy during the compression process. Motivated by these challenges, we propose a dynamic attention-aware approach for task-agnostic prompt compression (DAC). This approach effectively integrates entropy and attention information, dynamically sensing entropy shifts during compression to achieve fine-grained prompt compression. Extensive experiments across various domains, including LongBench, GSM8K, and BBH, show that DAC consistently yields robust and substantial improvements across a diverse range of tasks and LLMs, offering compelling evidence of its efficacy.

IAM: Efficient Inference through Attention Mapping between Different-scale LLMs

Yi Zhao, Zuchao Li, hai zhao

LLMs encounter significant challenges in resource consumption nowadays, especially with long contexts. Despite extensive efforts dedicate to enhancing inference efficiency, these methods primarily exploit internal sparsity within the models, without leveraging external information for optimization. We identify the high similarity of attention matrices across different-scale LLMs, which offers a novel perspective for optimization. We first conduct a comprehensive analysis of how to measure similarity, how to select mapping Layers and whether mapping is consistency. Based on these insights, we introduce the IAM framework, which achieves dual benefits of accelerated attention computation and reduced KV cache usage by performing attention mapping between small and large LLMs. Our experimental results demonstrate that IAM can accelerate prefill by 15% and reduce KV cache usage by 22.1% without appreciably sacrificing performance. Experiments on different series of models show the generalizability of IAM. Importantly, it is also orthogonal to many existing KV cache optimization methods, making it a versatile addition to the current toolkit for enhancing LLM efficiency.

SoftCoT: Soft Chain-of-Thought for Efficient Reasoning with LLMs

Yige Xu, Xu Guo, Zhiwei Zeng, Chunyan Miao

Chain-of-Thought (CoT) reasoning enables Large Language Models (LLMs) to solve complex reasoning tasks by generating intermediate reasoning steps. However, most existing approaches focus on hard token decoding, which constrains reasoning within the discrete vocabulary space and may not always be optimal. While recent efforts explore continuous-space reasoning, they often require full-model fine-tuning and suffer from catastrophic forgetting, limiting their applicability to state-of-the-art LLMs that already perform well in zero-shot settings with a proper instruction. To address this challenge, we propose a novel approach for continuous-space reasoning that does not require modifying the LLM. Specifically, we employ a lightweight fixed assistant model to speculatively generate instance-specific soft thought tokens as the initial chain of thoughts, which are then mapped into the LLM's representation space via a trainable projection module. Experimental results on five reasoning benchmarks demonstrate that our method enhances LLM reasoning performance through supervised, parameter-efficient fine-tuning. Source code is available at https://github.com/xuyige/SoftCoT.

Beyond Similarity: A Gradient-based Graph Method for Instruction Tuning Data Selection

Yang Zhao, Li Du, Xiao Ding, Yangou Ouyang, Hepeng Wang, Kai Xiong, Jinglong Gao, Zhouhao Sun, Dongliang Xu, Qing Yang, Dongchen Li, Bing Qin, Ting Liu

Large language models (LLMs) have shown great potential across various industries due to their remarkable ability to generalize through instruction tuning. However, the limited availability of domain-specific data significantly hampers their performance on specialized tasks. While existing methods primarily focus on selecting training data from general datasets that are similar to the target domain, they often fail to consider the joint distribution of instructions, resulting in inefficient learning and suboptimal knowledge transfer. To address these challenges, we introduce **G2IS** (**G**radient-based **G**raph **I**nstruction **S**election), a novel method that constructs a mixed gradient-based instruction graph to capture the joint distribution and interdependencies among instructions. By accounting for the relationships between instructions, G2IS improves domain adaptation efficiency. Additionally, we propose a gradient walk algorithm to refine the data selection process, enhancing both training effectiveness and efficiency. Our experiments demonstrate that G2IS outperforms traditional methods across various domain adaptation tasks, yielding significant performance gains, particularly in complex, data-scarce scenarios. These results underscore the potential of G2IS in advancing the development of large, domain-specific models.

500xCompressor: Generalized Prompt Compression for Large Language Models

Zongqian Li, Yixuan Su, Nigel Collier

Prompt compression is important for large language models (LLMs) to increase inference speed, reduce costs, and improve user experience. However, current methods face challenges such as low compression ratios and potential training-test overlap during evaluation. To address these issues, we propose 500xCompressor, a method that compresses natural language contexts into a minimum of one special token and demonstrates strong generalization ability. The 500xCompressor introduces approximately 0.3% additional parameters and achieves compression ratios ranging from 6x to 500x, achieving 27-90% reduction in calculations and 55-83% memory savings when generating 100-400 tokens for new and reused prompts at 500x compression, while retaining 70-74% (F1) and 77-84% (Exact Match) of the LLM capabilities compared to using non-compressed prompts. It is designed to compress any text, answer various types of questions, and can be utilized by the original LLM without requiring fine-tuning. Initially, 500xCompressor was pretrained on the ArxivCorpus, followed by fine-tuning on the ArxivQA dataset, and subsequently evaluated on strictly unseen and cross-domain question answering (QA) datasets. This study shows that KV values outperform embeddings in preserving information at high compression ratios. The highly compressive nature of natural language prompts, even for detailed information, suggests potential for future applications and the development of a new LLM language.

TigerLLM - A Family of Bangla Large Language Models

Nishat Raihan, Marcos Zampieri

The development of Large Language Models (LLMs) remains heavily skewed towards English and a few other high-resource languages. This linguistic disparity is particularly evident for Bangla - the 5th most spoken language. A few initiatives attempted to create open-source Bangla LLMs with performance still behind high-resource languages and limited reproducibility. To address this gap, we introduce TigerLLM - a family of Bangla LLMs. Our results demonstrate that these models surpass all open-source alternatives and also outperform larger proprietary models like GPT3.5 across standard benchmarks, establishing TigerLLM as the new baseline for future Bangla language modeling.

Scaling Laws and Efficient Inference for Ternary Language Models

Tejas Vaidhya, Ayush Kaushal, Vineet Jain, Francis Couture-Harpin, Prashant Shishodia, Majid Behbahani, Irina Rish, Yuriy Nevmyvaka

Large language models (LLMs) are increasingly used across research and industry applications, yet their inference efficiency remains a significant challenge. As the computational power of modern GPU architectures continuously improves, their memory bandwidth and capacity have not scaled proportionally, creating a critical bottleneck during inference. To address this, we investigate ternary language models (TriLMs) that employ quantization-aware training to significantly reduce memory requirements. We first analyze the scalability of TriLMs by conducting a scaling law analysis, revealing that TriLMs benefit more from increasing training data than from scaling model parameters. Based on this observation, we introduce TriTera, an open suite of TriLMs trained on up to 1.2 trillion tokens, demonstrating sustained performance gains at scale. Furthermore, to improve inference efficiency, we propose novel 2-bit and 1.6-bit packing schemes for ternary weights, which demonstrate accelerated inference across various CPU architectures. Building on the 2-bit packing, we develop a GPU kernel called TriRun that accelerates end-to-end model inference by up to 5 \times compared to floating-point baselines. To encourage further exploration and development of TriLMs, we will release the TriTera suite and TriRun inference kernels. Overall, our work lays the foundation for building and deploying efficient LLMs, providing a valuable resource for the research community.

StitchLLM: Serving LLMs, One Block at a Time

Bodun Hu, Shuozhe Li, Saurabh Agarwal, Myungjin Lee, Akshay Jajoo, Jiamin Li, Le Xu, Geon-Woo Kim, Donghyun Kim, Hong Xu, Amy Zhang, Aditya Akella

The rapid evolution of large language models (LLMs) has revolutionized natural language processing (NLP) tasks such as text generation, translation, and comprehension. However, the increasing computational demands and inference costs of these models present significant challenges. This study investigates the dynamic and efficient utilization of pre-trained weights from open-sourced LLMs of varying parameter sizes to achieve an optimal balance between computational efficiency and task performance. Drawing inspiration from the dual-process theory of human cognition, we introduce StitchLLM: a dynamic model routing framework that employs a powerful bottom model to process all queries, and uses a lightweight routing mechanism to allocate computational resources appropriately. Our novel framework optimizes efficiency and maintains performance, leveraging a trainable stitching layer for seamless integration of decoder layers across different LLMs. Experimental results demonstrate that StitchLLM improves system throughput while minimizing performance degradation, offering a flexible solution for deploying LLMs in resource-constrained settings.

Leveraging Self-Attention for Input-Dependent Soft Prompting in LLMs

Ananth Muppidi, Abhilash Nandy, Sambaran Bandyopadhyay

The performance of large language models in domain-specific tasks necessitates fine-tuning, which is computationally expensive and technically challenging. This paper focuses on parameter-efficient fine-tuning using soft prompting, a promising approach that adapts pre-trained models to downstream tasks by learning a small set of parameters. We propose a novel Input Dependent Soft Prompting technique with a self-Attention Mechanism (ID-SPAM) that generates soft prompts based on the input tokens and attends different tokens with varying importance. Our method is simple and efficient, keeping the number of trainable parameters small. We show the merits of the proposed approach compared to state-of-the-art techniques on various tasks and show the improved zero shot domain transfer capability.

DRPruning: Efficient Large Language Model Pruning through Distributionally Robust Optimization

Hexuan Deng, Wenxiang Jiao, Xuebo Liu, Jing Li, Min Zhang, Zhaopeng Tu

Large language models (LLMs) deliver impressive results but face challenges from increasing model sizes and computational costs. Structured pruning reduces model size and speeds up inference but often causes uneven degradation across domains, leading to biased performance. To address this, we propose *DRPruning*, a method that dynamically adjusts the data distribution during training to restore balanced performance across heterogeneous and multi-tasking data. Experiments in monolingual and multilingual settings show that DRPruning surpasses similarly sized models in both pruning and continued pretraining over perplexity, downstream tasks, and instruction tuning. Further analysis demonstrates the robustness of DRPruning towards various domains and distribution shifts. Furthermore, DRPruning can determine optimal reference losses and data ratios automatically, suggesting potential for broader applications. Code and scripts are available at https://github.com/hexuandeng/DRPruning.

A Parameter-Efficient and Fine-Grained Prompt Learning for Vision-Language Models

Yongbin Guo, Shuzhen Li, zhulin liu, Tong Zhang, C.L.Philip Chen

Current vision-language models (VLMs) understand complex vision-text tasks by extracting overall semantic information from large-scale cross-modal associations. However, extracting from large-scale cross-modal associations often smooths out semantic details and requires large computations, limiting multimodal fine-grained understanding performance and efficiency. To address this issue, this paper proposes a detail-oriented prompt learning (DoPL) method for vision-language models to implement fine-grained multi-modal semantic alignment with merely 0.25M trainable parameters. According to the low-entropy information concentration theory, DoPL explores shared interest tokens from text-vision correlations and transforms them into alignment weights to enhance text prompt and vision prompt via detail-oriented prompt generation. It effectively guides the current frozen layer to extract fine-grained text-vision alignment cues. Furthermore, DoPL constructs detail-oriented prompt generation for each frozen layer to implement layer-by-layer localization of fine-grained semantic alignment, achieving precise understanding in complex vision-text tasks. DoPL performs well in parameter-efficient fine-grained semantic alignment with only 0.12% tunable parameters for vision-language models. The state-of-the-art results over the previous parameter-efficient fine-tuning methods and full fine-tuning approaches on six benchmarks demonstrate the effectiveness and efficiency of DoPL in complex multi-modal tasks.

Efficient Many-Shot In-Context Learning with Dynamic Block-Sparse Attention

Emily Xiao, Chin-Jou Li, Yilin Zhang, Graham Neubig, Amanda Bertsch

Many-shot in-context learning has recently shown promise as an alternative to finetuning, with the major advantage that the same model can be served for multiple tasks. However, this shifts the computational burden from training-time to inference-time, making deployment of many-shot ICL challenging to justify in-practice. This cost is further increased if a custom demonstration set is retrieved for each inference example. We present Dynamic Block-Sparse Attention, an optimized method for retrieval-based many-shot in-context learning. By combining carefully designed block-sparse attention and retrieval of cached groups of demonstrations, we achieve comparable per-example latency to finetuning while maintaining on average >95% of the best method's accuracy across strong ICL and finetuning baselines. We hope that this will further enable the deployment of many-shot ICL at scale.

Cross-Lingual Representation Alignment Through Contrastive Image-Caption Tuning

Nathaniel Krasner, Nicholas Lanuzo, Antonios Anastasopoulos

Multilingual alignment of sentence representations has mostly required bitexts to bridge the gap between languages. We investigate whether visual information can bridge this gap instead. Image caption datasets are very easy to create without requiring multilingual expertise, so this offers a more efficient alternative for low-resource languages. We find that multilingual image-caption alignment can implicitly align the text representations between languages, languages unseen by the encoder in pretraining can be incorporated into this alignment post-hoc, and these aligned representations are usable for cross-lingual Natural Language Understanding (NLU) and bitext retrieval.

Squeezed Attention: Fast Fixed Context Processing for Long Context Length LLM Applications

Coleman Richard Charles Hooper, Sehoon Kim, Hiva Mohammadzadeh, Monishwaran Maheswaran, Sebastian Zhao, June Paik, Michael W. Mahoney, Kurt Keutzer, Amir Gholami

Emerging Large Language Model (LLM) applications require long input context in order to perform complex tasks like document analysis and code generation. For these long context length applications, the length of the input prompt poses a significant challenge in terms of inference efficiency since the inference costs increase linearly with sequence length. However, for many of these applications, much of the context in the prompt is fixed across different user inputs, thereby providing the opportunity to perform offline optimizations in order to process user inputs quickly, as they are received. We propose Squeezed Attention to accelerate LLM applications where a large portion of the input context is fixed. We first leverage K-means clustering offline to group the keys for the fixed context based on semantic similarity and represent each cluster with a single centroid value. During inference, we compare query tokens from the user input with the centroids to predict which keys from the fixed context are semantically relevant, and then compute exact attention using only the important keys, thereby reducing bandwidth and computational costs. We also present a hierarchical version of our algorithm which can reduce the complexity of attention from linear to logarithmic with respect to the fixed context length. We evaluate our method on various long-context benchmarks including LongBench, where it achieves a 3.1\times reduction in KV budget with no noticeable accuracy loss and up to an 8\times reduction with only a 0.5 point accuracy gap for the LLaMA-2-7B-32K, LWM-Text-Chat-1M, and Longchat-7B-v1.5-32K models. Futhermore, we implement kernels for centroid comparison and sparse FlashAttention with important keys, achieving more than 4\times speedups during both the prefill and generation phases for long-context inference. Our code is available at https://github.com/SqueezeAILab/SqueezedAttention.

Session 10: Efficient/Low-Resource Methods for NLP (Tuesday 29 July, 16:00-17:30, Hall 5X)

Token Level Routing Inference System for Edge Devices

Jianshu She, Wenhao Zheng, Zhengzhong Liu, Hongyi Wang, Eric P. Xing, Huaxiu Yao, Qirong Ho

The computational complexity of large language model (LLM) inference significantly constrains their deployment efficiency on edge devices. In contrast, small language models offer faster decoding and lower resource consumption but often suffer from degraded response quality and heightened susceptibility to hallucinations. To address this trade-off, collaborative decoding, in which a large model assists in generating critical tokens, has emerged as a promising solution. This paradigm leverages the strengths of both model types by enabling high-quality inference through selective intervention of the large model, while maintaining the speed and efficiency of the smaller model. In this work, we present a novel collaborative decoding inference system that allows small models to perform on-device inference while selectively consulting a cloud-based large model for critical token generation. Remarkably, the system achieves a 60\% performance gain on CommonsenseQA using only a 0.5B model on an M1 MacBook, with under 7\% of tokens generation uploaded to the large model in the cloud.

SlimLM: An Efficient Small Language Model for On-Device Document Assistance

Thang M. Pham, Phat T. Nguyen, Seunghyun Yoon, Viet Dac Lai, Franck Dernoncourt, Trung Bui

While small language models (SLMs) show promises for mobile deployment, their real world performance and applications on smartphones remain underexplored. We present SlimLM, a series of SLMs optimized for document assistance tasks on mobile devices. Through extensive experiments on a Samsung Galaxy S24, we identify the sweet spot between model size (ranging from 125M to 8B parameters), context length, and inference time for efficient on-device processing. SlimLM is pretrained on SlimPajama-627B and fine-tuned on DocAssist, our constructed dataset for summarization, question answering, and suggestion tasks. Our smallest model demonstrates efficient performance on S24, while larger variants offer enhanced capabilities within mobile constraints. We evaluate SlimLM against existing SLMs, showing comparable or superior performance and offering a benchmark for future research in on-device language models. We provide an Android application allowing users to experience SlimLM's document assistance capabilities, offering valuable insights for mobile developers, researchers, and companies seeking privacy-preserving on-device alternatives to server-based language models.

Session 10: Ethics, Bias, and Fairness (Tuesday 29 July, 16:00-17:30, Hall 4/5)

Sampling-based Pseudo-Likelihood for Membership Inference Attacks

Masahiro Kaneko, Youmi Ma, Yuki Wata, Naoaki Okazaki

Large Language Models (LLMs) are trained on large-scale web data, which makes it difficult to grasp the contribution of each text. This poses the risk of leaking inappropriate data such as benchmarks, personal information, and copyrighted texts in the training data. Membership Inference Attacks (MIA), which determine whether a given text is included in the model's training data, have been attracting attention. Previous studies of MIAs revealed that likelihood-based classification is effective for detecting leaks in LLMs. However, the existing likelihood-based methods cannot be applied to some proprietary models like ChatGPT or Claude 3 because the likelihood for input text is unavailable to the user. In this study, we propose a Sampling-based Pseudo-Likelihood (SPL) method for MIA (SaMIA) that calculates SPL using only the text generated by an LLM to detect leaks. The SaMIA treats the target text as the reference text and multiple outputs from the LLM as text samples, calculates the degree of n-gram match as SPL, and determines the membership of the text in the training data. Even without likelihoods, SaMIA performed on par with existing likelihood-based methods.

Social Bias Benchmark for Generation: A Comparison of Generation and QA-Based Evaluations

Jiho Jin, Woosung Kang, Junho Myung, Alice Oh

Measuring social bias in large language models (LLMs) is crucial, but existing bias evaluation methods struggle to assess bias in long-form generation. We propose a Bias Benchmark for Generation (BBG), an adaptation of the Bias Benchmark for QA (BBQ), designed to evaluate social bias in long-form generation by having LLMs generate continuations of story prompts. Building our benchmark in English and Korean, we measure the probability of neutral and biased generations across ten LLMs. We also compare our long-form story generation evaluation results with multiple-choice BBQ evaluation, showing that the two approaches produce inconsistent results.

Awes, Laws, and Flaws From Today's LLM Research

Adrian de Wynter

We perform a critical examination of the scientific methodology behind contemporary large language model (LLM) research. For this we assess over 2,000 research works released between 2020 and 2024 based on criteria typical of what is considered good research (e.g. presence of statistical tests and reproducibility), and cross-validate it with arguments that are at the centre of controversy (e.g., claims of emergent behaviour). We find multiple trends, such as declines in ethics disclaimers, a rise of LLMs as evaluators, and an increase on claims of LLM reasoning abilities without leveraging human evaluation. We note that conference checklists are effective at curtailing some of these issues, but balancing velocity and rigour in research cannot solely rely on these. We tie all these findings to findings from recent meta-reviews and extend recommendations on how to address what does, does not, and should work in LLM research.

Exploiting Instruction-Following Retrievers for Malicious Information Retrieval

Parishad BehnamGhader, Nicholas Meade, Siva Reddy

While instruction-following retrievers have seen widespread adoption alongside LLMs in real-world applications, little work has investigated the safety risks surrounding their increasing search capabilities. In this work, we empirically study the ability of retrievers to fulfill malicious queries, both when used directly and when used in a retrieval augmented generation-based setup. Concretely, we investigate six leading retrievers, including NV-Embed and LLM2Vec, and find that given malicious requests, most retrievers can (for >50% of queries) select relevant harmful passages. We further uncover an emerging risk with instruction-following retrievers, where highly relevant harmful information can be surfaced through fine-grained instructions. Finally, we show that even safety-aligned LLMs, such as Llama3, can produce harmful responses when provided with harmful retrieved passages in-context. In summary, our findings underscore the malicious misuse risks associated with increasing retriever capability.

Exploring Multimodal Challenges in Toxic Chinese Detection: Taxonomy, Benchmark, and Findings

Shujian Yang, Shiyao Cui, Haicheng Wang, Tianwei Zhang, Minlie Huang, Jialiang LU, Han Qiu

Detecting toxic content using language models is important but challenging. While large language models (LLMs) have demonstrated strong performance in understanding Chinese, recent studies show that simple character substitutions in toxic Chinese text can easily confuse the state-of-the-art (SOTA) LLMs. In this paper, we highlight the multimodal nature of Chinese language as a key challenge for deploying LLMs in toxic Chinese detection. First, we propose a taxonomy of 3 perturbation strategies and 8 specific approaches in toxic Chinese content. Then, we curate a dataset based on this taxonomy, and benchmark 9 SOTA LLMs (from both the US and China) to assess if they can detect perturbed toxic Chinese text. Additionally, we explore cost-effective enhancement solutions like in-context learning (ICL) and supervised fine-tuning (SFT). Our results reveal two important findings. (1) LLMs are less capable of detecting perturbed multimodal Chinese toxic contents. (2) ICL or SFT with a small number of perturbed examples may cause the LLMs "overcorrect": misidentify many normal Chinese contents as toxic.

The Impact of Name Age Perception on Job Recommendations in LLMs

Mahammed Kamruzzaman, Gene Louis Kim

Names often carry generational connotations, with certain names stereotypically associated with younger or older age groups. This study examines implicit age-related name bias in LLMs used for job recommendations. Analyzing six LLMs and 117 American names categorized by perceived age across 30 occupations, we find systematic bias: older-sounding names are favored for senior roles, while younger-sounding names are linked to youth-dominant jobs, reinforcing generational stereotypes. We also find that this bias is based on perceived rather than real ages associated with the names.

FedDQC: Data Quality Control in Federated Instruction-tuning of Large Language Models

Yaxin Du, Rui Ye, Fengting Yuchi, Wanru Zhao, Jingjing Qu, Yanfeng Wang, Siheng Chen

Federated Learning (FL) enables privacy-preserving collaborative instruction tuning of large language models (LLMs) by leveraging massively distributed data. However, the decentralized nature of FL exacerbates data quality challenges, as local clients lack global visibility to filter noisy or low-quality samples before training. To resolve this issue, we propose FedDQC, a novel federated instruction tuning framework with dynamic data quality control. Our approach introduces two key innovations. First, we propose instruction-response alignment (IRA)—an efficient client-side metric for quality evaluation requiring only low-cost inference. We validate that higher-IRA data corresponds to more relevant and easier-to-learn question-answer pairs. Second, mirroring the human easy-to-hard knowledge acquisition process, we design a quality-aware hierarchical FL training framework, where the LLM is progressively fine-tuned from high- to low-IRA data in a collaborative manner. The framework also supports adaptive data quality assessment at each hierarchy, enabling dynamic adjustments throughout the training process. Extensive experiments on synthetic and real-world datasets show that our method significantly improves LLM performance on mixed-quality data in FL.

Analyzing Political Bias in LLMs via Target-Oriented Sentiment Classification

Akram Elbouanani, Evan Dufraisse, Adrian Popescu

Political biases encoded by LLMs might have detrimental effects on downstream applications. Existing bias analysis methods rely on small-size intermediate tasks (questionnaire answering or political content generation) and rely on the LLMs themselves for analysis, thus propagating bias. We propose a new approach leveraging the observation that LLM sentiment predictions vary with the target entity in the same sentence. We define an entropy-based inconsistency metric to encode this prediction variability. We insert 1319 demographically and politically diverse politician names in 450 political sentences and predict target-oriented sentiment using seven models in six widely spoken languages. We observe inconsistencies in all tested combinations and aggregate them in a statistically robust analysis at different granularity levels. We observe positive and negative bias toward left and far-right politicians and positive correlations between politicians with similar alignment. Bias intensity is higher for Western languages than for others. Larger models exhibit stronger and more consistent biases and reduce discrepancies between similar languages. We partially mitigate LLM unreliability in target-oriented sentiment classification (TSC) by replacing politician names with fictional but plausible counterparts. The complete code, the data, and all analyses will be made public to enable reproducibility.

Navigating the Political Compass: Evaluating Multilingual LLMs across Languages and Nationalities

Chadi Helwe, Oana Balalau, Davide Ceolin

Large Language Models (LLMs) have become ubiquitous in today's technological landscape, boasting a plethora of applications, and even endangering human jobs in complex and creative fields. One such field is journalism: LLMs are being used for summarization, generation and even fact-checking. However, in today's political landscape, LLMs could accentuate tensions if they exhibit political bias. In this work, we evaluate the political bias of the most used 15 multilingual LLMs via the Political Compass Test. We test different scenarios, where we vary the language of the prompt, while also assigning a nationality to the model. We evaluate models on the 50 most populous countries and their official languages. Our results indicate that language has a strong influence on the political ideology displayed by a model. In addition, smaller models tend to display a more stable political ideology, i.e. ideology that is less affected by variations in the prompt.

A Survey on Proactive Defense Strategies Against Misinformation in Large Language Models

Shuliang Liu, Hongyi Liu, Aiwei Liu, Duan Bingchen, Zheng Qi, Yibo Yan, He GENG, Peijie Jiang, Jia Liu, Xuming Hu

The widespread deployment of large language models (LLMs) across critical domains has amplified the societal risks posed by algorithmically generated misinformation. Unlike traditional false content, LLM-generated misinformation can be self-reinforcing, highly plausible, and capable of rapid propagation across multiple languages, which traditional detection methods fail to mitigate effectively. This paper introduces a proactive defense paradigm, shifting from passive post hoc detection to anticipatory mitigation strategies. We propose a Three Pillars framework: (1) Knowledge Credibility, fortifying the integrity of training and deployed data; (2) Inference Reliability, embedding self-corrective mechanisms during reasoning; and (3) Input Robustness, enhancing the resilience of model interfaces against adversarial attacks. Through a comprehensive survey of existing techniques and a comparative meta-analysis, we demonstrate that proactive defense strategies offer up to 63% improvement over conventional methods in misinformation prevention, despite non-trivial computational overhead and generalization challenges. We argue that future research should focus on co-designing robust knowledge foundations, reasoning certification, and attack-resistant interfaces to ensure LLMs can effectively counter misinformation across varied domains.

Are LLMs Rational Investors? A Study on the Financial Bias in LLMs

Yuhang Zhou, Yuchen Ni, Zhiheng Xi, Zhangyue Yin, Yu He, Gan Yunhui, Xiang Liu, Zhang Jian, Sen Liu, Xipeng Qiu, Yixin Cao, Guangnan Ye, Hongfeng Chai

Large language models (LLMs) excel in natural language generation but also exhibit biases, particularly in gender, race, and religion, which can be amplified with widespread use. However, research on biases in specific domains, such as finance, remains limited. To address this gap, we conducted a comprehensive evaluation of 23 leading LLMs and found varying degrees of financial bias, including more pronounced biases in financial-specific LLMs (FinLLMs). In response, we propose the Financial Bias Indicators (FBI) framework, which includes components like the Bias Unveiler, Bias Detective, Bias Tracker, and Bias Antidote, designed to identify, detect, analyze, and mitigate financial biases. Our analysis explores the root causes of these biases and introduces a debiasing method based on financial causal knowledge, alongside three other debiasing techniques. For the most biased model, we successfully reduced bias by 68% according to key metrics. This study advances our understanding of LLM biases in finance and highlights the need for greater scrutiny in their application within this critical domain.

Stereotype or Personalization? User Identity Biases Chatbot Recommendations

Anjali Kantharuban, Jeremiah Milbauer, Maarten Sap, Emma Strubell, Graham Neubig

While personalized recommendations are often desired by users, it can be difficult in practice to distinguish cases of bias from cases of personalization: we find that models generate racially stereotypical recommendations regardless of whether the user revealed their identity intentionally through explicit indications or unintentionally through implicit cues. We demonstrate that when people use large language models (LLMs) to generate recommendations, the LLMs produce responses that reflect both what the user wants and who the user is. We argue that chatbots ought to transparently indicate when recommendations are influenced by a user's revealed identity characteristics, but observe that they currently fail to do so. Our experiments show that even though a user's revealed identity significantly influences model recommendations ($p < 0.001$), model responses obfuscate this fact in response to user queries. This bias and lack of transparency occurs consistently across multiple popular consumer LLMs and for four American racial groups.

"You are Beautiful, Body Image Stereotypes are Ugly!" BIStereo: A Benchmark to Measure Body Image Stereotypes in Language Models

Narjis Asad, Nihar Ranjan Sahoo, Rudra Murthy, Swaprava Nath, Pushpak Bhattacharyya

While a few high-quality bias benchmark datasets exist to address stereotypes in Language Models (LMs), a notable lack of focus remains on body image stereotypes. To bridge this gap, we propose \textbf{BIStereo}, a suite to uncover LMs' biases towards people of certain physical appearance characteristics, namely, \textit{skin complexion, body shape, height, attire,} and a \textit{miscellaneous category} including \textit{hair texture, eye color, and more}. Our dataset comprises 40k sentence pairs designed to assess LMs' biased preference for certain body types. We further include 60k premise-hypothesis pairs designed to comprehensively assess LMs' preference for fair skin tone. Additionally, we curate 553 tuples consisting of a \textit{body image descriptor, gender, and a stereotypical attribute}, validated by a diverse pool of annotators for physical appearance stereotypes. We propose a metric, \textbf{TriSentBias}, that captures the biased preferences of LMs towards a certain body type over others. Using \textbf{BIStereo}, we assess the presence of body image biases in ten different language models, revealing significant biases in models Muril, XLMR, Llama3, and Gemma. We further evaluate the LMs through downstream NLI and Analogy tasks. Our NLI experiments highlight notable patterns in the LMs that align with the well-documented cognitive bias in humans known as \textbf{\textit{the Halo Effect}}. All scripts and datasets from this study will be publicly available.

Shiny Inputs, Shiny Outputs? Unveiling the Halo Effect in MultiModal LLMs

Kyusik Kim, Jeongwoo Ryu, Hyeonseok Jeon, Bongwon Suh

This study investigates the halo effect in AI-driven hiring evaluations using Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs). Through experiments with hypothetical job applications, we examined how these models' evaluations are influenced by non-job-related information including extracurricular activities, social media images, and video interview backgrounds. By analyzing models' responses to Likert-scale questions across different competency dimensions, we found that AI models exhibit significant halo effects, particularly in image-based evaluations, while text-based assessments showed more resistance to bias. The findings demonstrate that supplementary multimodal information can substantially influence AI hiring decisions, highlighting potential risks in AI-based recruitment systems.

Large Language Models Still Exhibit Bias in Long Text

Wonje Jeung, Dongjae Jeon, Ashkan Yousefpour, Jonghyun Choi

Existing fairness benchmarks for large language models (LLMs) primarily focus on simple tasks, such as multiple-choice questions, overlooking biases that may arise in more complex scenarios like long-text generation. To address this gap, we introduce the Long Text Fairness Test (LTF-TEST), a framework that evaluates biases in LLMs through essay-style prompts. LTF-TEST covers 14 topics and 10 demographic axes, including gender and race, resulting in 11,948 samples. By assessing both model responses and the reasoning behind them, LTF-TEST uncovers subtle biases that are difficult to detect in simple responses. In our evaluation of five recent LLMs, including GPT-4o and LLaMA3, we identify two key patterns of bias. First, these models frequently favor certain demographic groups in their responses. Second, they show excessive sensitivity toward traditionally disadvantaged groups, often providing overly protective responses while neglecting others. To mitigate these biases, we propose REGARD-FT, a finetuning approach that pairs biased prompts with neutral responses. REGARD-FT reduces gender bias by 34.6% and improves performance by 1.4 percentage points on the BBQ benchmark, offering a promising approach to addressing biases in long-text generation tasks.

7 Points to Tsinghua but 10 Points to ? Assessing Large Language Models in Agentic Multilingual National Bias

Qianying Liu, Katrina Qiyao Wang, Fei Cheng, Sadao Kurohashi

Large Language Models have garnered significant attention for their capabilities in multilingual natural language processing, while studies on risks associated with cross biases are limited to immediate context preferences. Cross-language disparities in reasoning-based recommendations remain largely unexplored, with a lack of even descriptive analysis. This study is the first to address this gap. We test LLM's applicability and capability in providing personalized advice across three key scenarios: university applications, travel, and relocation. We investigate multilingual bias in state-of-the-art LLMs by analyzing their responses to decision-making tasks across multiple languages. We quantify bias in model-generated scores and assess the impact of demographic factors and reasoning strategies (e.g., Chain-of-Thought prompting) on bias patterns. Our findings reveal significant biases in both the scores and the reasoning structure of non-English languages. We also draw future implications for improving multilingual alignment in AI systems.

LLMs are Biased Evaluators But Not Biased for Retrieval Augmented Generation

Yen-Shan Chen, Jin Jing, Peng-Ting Kuo, Chao-Wei Huang, Yun-Nung Chen

Recent studies have demonstrated that large language models (LLMs) exhibit significant biases in evaluation tasks, particularly in preferentially rating and favoring self-generated content. However, the extent to which this bias manifests in fact-oriented tasks, especially within retrieval-augmented generation (RAG) frameworks—where keyword extraction and factual accuracy take precedence over stylistic elements—remains unclear. Our study addresses this knowledge gap by simulating two critical phases of the RAG framework. In the first phase, LLMs evaluated human-authored and model-generated passages, emulating the pointwise reranking phase. The second phase involves conducting pairwise reading comprehension tests to simulate the generation phase. Contrary to previous findings indicating a self-preference in rating tasks, our results reveal no significant self-preference effect in RAG frameworks. Instead, we observe that factual accuracy significantly influences LLMs' output, even in the absence of prior knowledge. These findings are consistent among three common QA datasets (NQ, MARCO, TriviaQA Datasets) and 5 widely adopted language models (GPT-3.5, GPT-4o-mini, Gemini, LLaMA3, and Mistral). Our research contributes to the ongoing discourse on LLM biases and their implications for RAG-based system, offering insights that may inform the development of more robust and unbiased LLM systems.

The Impossibility of Fair LLMs

Jacy Reese Anthis, Kristian Lum, Michael Ekstrand, Avi Feller, Chenhao Tan

The rise of general-purpose artificial intelligence (AI) systems, particularly large language models (LLMs), has raised pressing moral questions about how to reduce bias and ensure fairness at scale. Researchers have documented a sort of "bias" in the significant correlations between demographics (e.g., race, gender) in LLM prompts and responses, but it remains unclear how LLM fairness could be evaluated with more rigorous definitions, such as group fairness or fair representations. We analyze a variety of technical fairness frameworks and find inherent challenges in each that make the development of a fair LLM intractable. We show that each framework either does not logically extend to the general-purpose AI context or is infeasible in practice, primarily due to the large amounts of unstructured training data and the many potential combinations of human populations, use cases, and sensitive attributes. These inherent challenges would persist for general-purpose AI, including LLMs, even if empirical challenges, such as limited participatory input and limited measurement methods, were overcome. Nonetheless, fairness will remain an important type of model evaluation, and there are still promising research directions, particularly the development of standards for the responsibility of LLM developers, context-specific evaluations, and methods of iterative, participatory, and AI-assisted evaluation that could scale fairness across the diverse contexts of modern human-AI interaction.

Delving into Multilingual Ethical Bias: The MSQAD with Statistical Hypothesis Tests for Large Language Models

Seunguk Yu, Juhwan Choi, YoungBin Kim

Despite the recent strides in large language models, studies have underscored the existence of social biases within these systems. In this paper, we delve into the validation and comparison of the ethical biases of LLMs concerning globally discussed and potentially sensitive topics, hypothesizing that these biases may arise from language-specific distinctions. Introducing the Multilingual Sensitive Questions \& Answers Dataset (**MSQAD**), we collected news articles from Human Rights Watch covering 17 topics, and generated socially sensitive questions along with corresponding responses in multiple languages. We scrutinized the biases of these responses across languages and topics, employing two statistical hypothesis tests. The results showed that the null hypotheses were rejected in most cases, indicating biases arising from cross-language differences. It demonstrates that ethical biases in responses are widespread across various languages, and notably, these biases were prevalent even among different LLMs. By making the proposed MSQAD openly available, we aim to facilitate future research endeavors focused on examining cross-language biases in LLMs and their variant models.

Evaluation of LLM Vulnerabilities to Being Misused for Personalized Disinformation Generation

Aneta Zugecova, Dominik Macko, Ivan Srba, Robert Moro, Jakub Kopál, Katarína Marcinčinová, Matúš Mesarčík

The capabilities of recent large language models (LLMs) to generate high-quality content indistinguishable by humans from human-written texts raises many concerns regarding their misuse. Previous research has shown that LLMs can be effectively misused for generating disinformation news articles following predefined narratives. Their capabilities to generate personalized (in various aspects) content have also been evaluated and mostly found usable. However, a combination of personalization and disinformation abilities of LLMs has not been comprehensively studied yet. Such a dangerous combination should trigger integrated safety filters of the LLMs, if there are some. This study fills this gap by evaluating vulnerabilities of recent open and closed LLMs, and their willingness to generate personalized disinformation news articles in English. We further explore whether the LLMs can reliably meta-evaluate the personalization quality and whether the personalization affects the generated-texts detectability. Our results demonstrate the need for stronger safety-filters and disclaimers, as those are not properly functioning in most of the evaluated LLMs. Additionally, our study revealed that the personalization actually reduces the safety-filter activations; thus effectively functioning as a jailbreak. Such behavior must be urgently addressed by LLM developers and service providers.

On the Mutual Influence of Gender and Occupation in LLM Representations

Haozhe An, Connor Baumler, Abhilasha Sancheti, Rachel Rudinger

We examine LLM representations of gender for first names in various occupational contexts to study how occupations and the gender perception of first names in LLMs influence each other mutually. We find that LLMs' first-name gender representations correlate with real-world gender statistics associated with the name, and are influenced by the co-occurrence of stereotypically feminine or masculine occupations. Additionally, we study the influence of first-name gender representations on LLMs in a downstream occupation prediction task and their potential as an internal metric to identify extrinsic model biases. While feminine first-name embeddings often raise the probabilities for female-dominated jobs (and vice versa for male-dominated jobs), reliably using these internal gender representations for bias detection remains challenging.

What Really Matters in Many-Shot Attacks? An Empirical Study of Long-Context Vulnerabilities in LLMs

Sangyeop Kim, Yohan Lee, Yongwoo Song, Kimin Lee

We investigate long-context vulnerabilities in Large Language Models (LLMs) through Many-Shot Jailbreaking (MSJ). Our experiments utilize context length of up to 128K tokens. Through comprehensive analysis with various many-shot attack settings with different instruction styles, shot density, topic, and format, we reveal that context length is the primary factor determining attack effectiveness. Critically, we find that successful attacks do not require carefully crafted harmful content. Even repetitive shots or random dummy text can circumvent model safety measures, suggesting fundamental limitations in long-context processing capabilities of LLMs. The safety behavior of well-aligned models becomes increasingly inconsistent with longer contexts. These findings highlight significant safety gaps in context expansion capabilities of LLMs, emphasizing the need for new safety mechanisms.

Quantification of Large Language Model Distillation

Sunbowen Lee, Junting Zhou, Chang Ao, Kaige Li, Xeron Du, Sirui He, Haihong Wu, Tianci Liu, Jiaheng Liu, Hamid Alinejad-Rokny, Min Yang, Yitao Liang, Zhoufutu Wen, Shiwen Ni

Model distillation is a fundamental technique in building large language models (LLMs), transferring knowledge from a teacher model to a student model. However, distillation can lead to model homogenization, reducing diversity among models and impairing their ability to robustly handle complex or novel tasks. These limitations underscore the need to systematically quantify the distillation process and its impact. In this work, we propose a framework to evaluate and quantify model distillation. Our method addresses two key aspects: (1) Identifying identity cognition contradictions to assess discrepancies in how models perceive and represent identity-related information, and (2) Analyzing multi-granularity response similarities across models to measure the extent of homogenization. Experimental results demonstrate two key insights: (1) Well-known closed-source and open-source LLMs usually exhibit high distillation degrees, except for Claude, Doubao, and Gemini. (2) Base LLMs show higher distillation degrees compared to aligned LLMs. By offering a systematic approach to improve the transparency of LLM data distillation, we call for LLMs with more independent development and more transparent technical reports to improve LLMs' robustness and safety. The code and data are available at https://github.com/Aegis1863/LLMs-Distillation-Quantification.

Mitigating Selection Bias with Node Pruning and Auxiliary Options

Hyeong Kyu Choi, Weijie Xu, Chi Xue, Stephanie Eckman, Chandan K. Reddy

Large language models (LLMs) often exhibit systematic preferences for certain answer choices when responding to multiple-choice questions—a behavior known as selection bias. This bias reduces the accuracy and reliability of LLM outputs, limiting their usefulness in decision-critical applications. While prior work has focused on adjusting model inputs or outputs to mitigate this issue, our work takes a fundamentally different approach by identifying and removing the internal sources of bias. We introduce two methods: Bias Node Pruning (BNP), which prunes parameters that contribute to selection bias, and Auxiliary Option Injection (AOI), which introduces an additional answer choice to reduce bias in both white-box and black-box settings. To address the shortcomings of existing evaluation metrics, we propose Choice Kullback-Leibler Divergence (CKLD), a new metric that captures distributional imbalances in model predictions. Experiments on three LLMs across multiple datasets demonstrate that our methods consistently improve answer accuracy while reducing selection bias, providing a robust solution for both open- and closed-source models.

Deontological Keyword Bias: The Impact of Modal Verbs on Normative Judgments of Language Models

Bumjin Park, Leejinsil, Jaesik Choi

Large language models (LLMs) are increasingly engaging in moral and ethical reasoning, where criteria for judgment are often unclear, even for humans. While LLM alignment studies cover many areas, one important yet underexplored area is how LLMs make judgments about obligations. This work reveals a strong tendency in LLMs to judge non-obligatory contexts as obligations when prompts are augmented with modal expressions such as \textit{must} or \textit{ought to}. We introduce this phenomenon as Deontological Keyword Bias (DKB). We find that LLMs judge over 90\% of commonsense scenarios as obligations when modal expressions are present. This tendency is consist across various LLM families, question types, and answer formats. To mitigate DKB, we propose a judgment strategy that integrates few-shot examples with reasoning prompts. This study sheds light on how modal expressions, as a form of linguistic framing, influence the normative decisions of LLMs and underscores the importance of addressing such biases to ensure judgment alignment.

VLSBench: Unveiling Visual Leakage in Multimodal Safety

Xuhao Hu, Dongrui Liu, Hao Li, Xuanjing Huang, Jing Shao

Safety concerns of Multimodal large language models (MLLMs) have gradually become an important problem in various applications. Surprisingly, previous works indicate a counterintuitive phenomenon that using textual unlearning to align MLLMs achieves comparable safety performances with MLLMs aligned with image-text pairs. To explain such a phenomenon, we discover a \textit{\textbf{V}isual \textbf{S}afety \textbf{I}nformation \textbf{L}eakage} (\textbf{VSIL}) problem in existing multimodal safety benchmarks, \textit{i.e.}, the potentially risky content in the image has been revealed in the textual query. Thus, MLLMs can easily refuse these sensitive image-text pairs according to textual queries only, leading to \textbf{unreliable cross-modality safety evaluation of MLLMs}. We also conduct a further comparison experiment between textual alignment and multimodal alignment to highlight this drawback. To this end, we construct \textit{\textbf{V}isual \textbf{L}eakless \textbf{S}afety \textbf{B}ench} (\textbf{VLSBench}) with 2.2k image-text pairs through an automated data pipeline. Experimental results indicate that VLSBench poses a significant challenge to both open-source and close-source MLLMs, \textit{i.e.}, LLaVA, Qwen2-VL and GPT-4o. Besides, we empirically compare textual and multimodal alignment methods on VLSBench and find that textual alignment is effective enough for multimodal safety scenarios with VSIL, while multimodal alignment is preferable for safety scenarios without VSIL.

Towards Geo-Culturally Grounded LLM Generations

Piyawat Lertvittayakumjorn, David Kinney, Vinodkumar Prabhakaran, Donald Martin Jr., Sunipa Dev

Generative large language models (LLMs) have demonstrated gaps in diverse cultural awareness across the globe. We investigate the effect of retrieval augmented generation and search-grounding techniques on LLMs' ability to display familiarity with various national cultures. Specifically, we compare the performance of standard LLMs, LLMs augmented with retrievals from a bespoke knowledge base (i.e., KB grounding), and LLMs augmented with retrievals from a web search (i.e., search grounding) on multiple cultural awareness benchmarks. We find that search grounding significantly improves the LLM performance on multiple-choice benchmarks that test propositional knowledge (e.g., cultural norms, artifacts, and institutions), while KB grounding's effectiveness is limited by inadequate knowledge base coverage and a suboptimal retriever. However, search grounding also increases the risk of stereotypical judgments by language models and fails to improve evaluators' judgments of cultural familiarity in a human evaluation with adequate statistical power. These results highlight the distinction between propositional cultural knowledge and open-ended cultural fluency when it comes to evaluating LLMs' cultural awareness.

PIG: Privacy Jailbreak Attack on LLMs via Gradient-based Iterative In-Context Optimization

Yidan Wang, Yanan Cao, Yubing Ren, Fang Fang, Zheng Lin, Binxing Fang

Large Language Models (LLMs) excel in various domains but pose inherent privacy risks. Existing methods to evaluate privacy leakage in LLMs often use memorized prefixes or simple instructions to extract data, both of which well-alignment models can easily block. Meanwhile, Jailbreak attacks bypass LLM safety mechanisms to generate harmful content, but their role in privacy scenarios remains underexplored. In this paper, we examine the effectiveness of jailbreak attacks in extracting sensitive information, bridging privacy leakage and jailbreak attacks in LLMs. Moreover, we propose PIG, a novel framework targeting Personally Identifiable Information (PII) and addressing the limitations of current jailbreak methods. Specifically, PIG identifies PII entities and their types in privacy queries, uses in-context learning to build a privacy context, and iteratively updates it with three gradient-based strategies to elicit target PII. We evaluate PIG and existing jailbreak methods using two privacy-related datasets. Experiments on four white-box and two black-box LLMs show that PIG outperforms baseline methods and achieves state-of-the-art (SoTA) results. The results underscore significant privacy risks in LLMs, emphasizing the need for stronger safeguards.

Towards Context-Robust LLMs: A Gated Representation Fine-tuning Approach

Shenglai Zeng, Pengfei He, Kai Guo, Tianqi Zheng, Hanqing Lu, Yue Xing, Hui Liu

Large Language Models (LLMs) enhanced with external contexts, such as through retrieval-augmented generation (RAG), often face challenges in handling imperfect evidence. They tend to over-rely on external knowledge, making them vulnerable to misleading and unhelpful contexts. To address this, we propose the concept of context-robust LLMs, which can effectively balance internal knowledge with external context, similar to human cognitive processes. Specifically, context-robust LLMs should rely on external context only when lacking internal knowledge, identify contradictions between internal and external knowledge, and disregard unhelpful contexts. To achieve this goal, we introduce Grft, a lightweight and plug-and-play gated representation fine-tuning approach. Grft consists of two key components: a gating mechanism to detect and filter problematic inputs, and low-rank representation adapters to adjust hidden representations. By training a lightweight intervention function with only 0.0004% of model size on fewer than 200 examples, Grft can effectively adapt LLMs towards context-robust behaviors.

Gradient-Adaptive Policy Optimization: Towards Multi-Objective Alignment of Large Language Models

Chengao Li, Hanyu Zhang, Yunkun Xu, Hongyan Xue, Xiang Ao, Qing He

Reinforcement Learning from Human Feedback (RLHF) has emerged as a powerful technique for aligning large language models (LLMs) with human preferences. However, effectively aligning LLMs with diverse human preferences remains a significant challenge, particularly when they are conflict. To address this issue, we frame human value alignment as a multi-objective optimization problem, aiming to maximize a set of potentially conflicting objectives. We introduce Gradient-Adaptive Policy Optimization (GAPO), a novel fine-tuning paradigm that employs multiple-gradient descent to align LLMs with diverse preference distributions. GAPO adaptively rescales the gradients for each objective to determine an update direction that optimally balances the trade-offs between objectives. Additionally, we introduce P-GAPO, which incorporates user preferences across different objectives and achieves Pareto solutions that better align with the user's specific needs.

Unmasking Style Sensitivity: A Causal Analysis of Bias Evaluation Instability in Large Language Models

Jiaxu Zhao, Meng Fang, Kun Zhang, Mykola Pechenizkiy

Natural language processing applications are increasingly prevalent, but social biases in their outputs remain a critical challenge. While various bias evaluation methods have been proposed, these assessments show unexpected instability when input texts undergo minor stylistic changes. This paper conducts a comprehensive analysis of how different style transformations impact bias evaluation results across multiple language models and bias types using causal inference techniques. Our findings reveal that formality transformations significantly affect bias scores, with informal style showing substantial bias reductions (up to 8.33% in LLaMA-2-13B). We identify appearance bias, sexual orientation bias, and religious bias as most susceptible to style changes, with variations exceeding 20%. Larger models demonstrate greater sensitivity to stylistic variations, with bias measurements fluctuating up to 3.1% more than in smaller models. These results highlight critical limitations in current bias evaluation methods and emphasize the need for reliable and fair assessments of language models.

One-Shot is Enough: Consolidating Multi-Turn Attacks into Efficient Single-Turn Prompts for LLMs

Junwoo Ha, Hyunjun Kim, Sangyoon Yu, Haon Park, Ashkan Yousefpour, Yuna Park, Suhyun Kim

We introduce a novel framework for consolidating multi-turn adversarial “jailbreak” prompts into single-turn queries, significantly reducing the manual overhead required for adversarial testing of large language models (LLMs). While multi-turn human jailbreaks have been shown to yield high attack success rates (ASRs), they demand considerable human effort and time. Our proposed Multi-turn-to-Single-turn (M2S) methods—Hyphenize, Numberize, and Pythonize—systematically reformat multi-turn dialogues into structured single-turn prompts. Despite eliminating iterative back-and-forth interactions, these reformatted prompts preserve and often enhance adversarial potency: in extensive evaluations on the Multi-turn Human Jailbreak (MHJ) dataset, M2S methods yield ASRs ranging from 70.6 % to 95.9 % across various state-of-the-art LLMs. Remarkably, our single-turn prompts outperform the original multi-turn attacks by up to 17.5 % in absolute ASR, while reducing token usage by more than half on average. Further analyses reveal that embedding malicious requests in enumerated or code-like structures exploits “contextual blindness,” undermining both native guardrails and external input-output safeguards. By consolidating multi-turn conversations into efficient single-turn prompts, our M2S framework provides a powerful tool for large-scale red-teaming and exposes critical vulnerabilities in contemporary LLM defenses. All code, data, and conversion prompts are available for reproducibility and further investigations: https://github.com/Junuha/M2S_DATA

Can't See the Forest for the Trees: Benchmarking Multimodal Safety Awareness for Multimodal LLMs

Wenxuan Wang, Xiaoyuan Liu, Kuiyi Gao, Jen-tse Huang, Youliang Yuan, Pinjia He, Shuai Wang, Zhaopeng Tu

Multimodal Large Language Models (MLLMs) have expanded the capabilities of traditional language models by enabling interaction through both text and images. However, ensuring the safety of these models remains a significant challenge, particularly in accurately identifying whether multimodal content is safe or unsafe—a capability we term \emph{safety awareness}. In this paper, we introduce MMSafeAware, the first comprehensive multimodal safety awareness benchmark designed to evaluate MLLMs across 29 safety scenarios with 1,500 carefully curated image-prompt pairs. MMSafeAware includes both unsafe and over-safety subsets to assess models' abilities to correctly identify unsafe content and avoid over-sensitivity that can hinder helpfulness. Evaluating nine widely used MLLMs using MMSafeAware reveals that current models are not sufficiently safe and often overly sensitive; for example, GPT-4V misclassifies 36.1% of unsafe inputs as safe and 59.9% of benign inputs as unsafe. We further explore three methods to improve safety awareness—prompting-based approaches, visual contrastive decoding, and vision-centric reasoning fine-tuning—but find that none achieve satisfactory performance. Our findings highlight the profound challenges in developing MLLMs with robust safety awareness, underscoring the need for further research in this area. All the code and data will be publicly available to facilitate future research.

Which Demographics do LLMs Default to During Annotation?

Johannes Schäfer, Aidan Combs, Christopher Bagdon, Jiahui Li, Nadine Probol, Lynn Greschner, Sean Papay, Yarik Menchaca Resendiz, Aswathy Velutharambath, Amelie Wuehrl, Sabine Weber, Roman Klinger

Demographics and cultural background of annotators influence the labels they assign in text annotation – for instance, an elderly woman might find it offensive to read a message addressed to a "bro", but a male teenager might find it appropriate. It is therefore important to acknowledge label variations to not under-represent members of a society. Two research directions developed out of this observation in the context of using large language models (LLM) for data annotations, namely (1) studying biases and inherent knowledge of LLMs and (2) injecting diversity in the output by manipulating the prompt with demographic information. We combine these two strands of research and ask the question to which demographics an LLM resorts to when no demographics is given. To answer this question, we evaluate which attributes of human annotators LLMs inherently mimic. Furthermore, we compare non-demographic conditioned prompts and placebo-conditioned prompts (e.g., "you are an annotator who lives in house number 5") to demographics-conditioned prompts ("You are a 45 year old man and an expert on politeness annotation. How do you rate {instance}"). We study these questions for politeness and offensiveness annotations on the POPQUORN data set, a corpus created in a controlled manner to investigate human label variations based on demographics which has not been used for LLM-based analyses so far. We observe notable influences related to gender, race, and age in demographic prompting, which contrasts with previous studies that found no such effects.

Insight Over Sight: Exploring the Vision-Knowledge Conflicts in Multimodal LLMs

Xiaoyuan Liu, Wenxuan Wang, Youliang Yuan, Jen-tse Huang, Qiuzhi Liu, Pinjia He, Zhaopeng Tu

This paper explores the problem of commonsense level vision-knowledge conflict in Multimodal Large Language Models (MLLMs), where visual information contradicts model's internal commonsense knowledge. To study this issue, we introduce an automated framework, augmented with human-in-the-loop quality control, to generate inputs designed to simulate and evaluate these conflicts in MLLMs. Using this framework, we have crafted a diagnostic benchmark consisting of 374 original images and 1,122 high-quality question-answer (QA) pairs. The benchmark covers two aspects of conflict and three question types, providing a thorough assessment tool. We apply this benchmark to assess the conflict-resolution capabilities of nine representative MLLMs from various model families. Our results indicate an evident over-reliance on parametric knowledge for approximately 20\% of all queries, especially among Yes-No and action-related problems. Based on these findings, we evaluate the effectiveness of existing approaches to mitigating the conflicts and compare them to our ``Focus-on-Vision'' prompting strategy. Despite some improvement, the vision-knowledge conflict remains unresolved and can be further scaled through our data construction framework. Our proposed framework, benchmark, and analysis contribute to the understanding and mitigation of vision-knowledge conflicts in MLLMs.

Guiding not Forcing: Enhancing the Transferability of Jailbreaking Attacks on LLMs via Removing Superfluous Constraints

Junxiao Yang, Zhexin Zhang, Shiyao Cui, Hongning Wang, Minlie Huang

Jailbreaking attacks can effectively induce unsafe behaviors in Large Language Models (LLMs); however, the transferability of these attacks across different models remains limited. This study aims to understand and enhance the transferability of gradient-based jailbreaking methods, which are among the standard approaches for attacking white-box models. Through a detailed analysis of the optimization process, we introduce a novel conceptual framework to elucidate transferability and identify superfluous constraints—specifically, the response pattern constraint and the token tail constraint—as significant barriers to improved transferability. Removing these unnecessary constraints substantially enhances the transferability and controllability of gradient-based attacks. Evaluated on Llama-3-8B-Instruct as the source model, our method increases the overall Transfer Attack Success Rate (T-ASR) across a set of target models with varying safety levels from 18.4\% to 50.3\%, while also improving the stability and controllability of jailbreak behaviors on both source and target models.

SEA: Low-Resource Safety Alignment for Multimodal Large Language Models via Synthetic Embeddings

Weikai Lu, Hao Peng, Huiping Zhuang, Cen Chen, Ziqian Zeng

Multimodal Large Language Models (MLLMs) have serious security vulnerabilities. While safety alignment using multimodal datasets consisting of text and data of additional modalities can effectively enhance MLLM's security, it is costly to construct these datasets. Existing low-resource security alignment methods, including textual alignment, have been found to struggle with the security risks posed by additional modalities. To address this, we propose Synthetic Embedding augmented safety Alignment (SEA), which optimizes embeddings of additional modality through gradient updates to expand textual datasets. This enables multimodal safety alignment training even when only textual data is available. Extensive experiments on image, video, and audio-based MLLMs demonstrate that SEA can synthesize a high-quality embedding on a single RTX3090 GPU within 24 seconds. SEA significantly improves the security of MLLMs when faced with threats from additional modalities. To assess the security risks introduced by video and audio, we also introduced a new benchmark called VA-SafetyBench. High attack success rates across multiple MLLMs validate its challenge. Our code and data will be available at https://github.com/ZeroNLP/SEA.

Estimating Privacy Leakage of Augmented Contextual Knowledge in Language Models

James Flemings, Bo Jiang, Wanrong Zhang, Zafar Takhirov, Murali Annavaram

Language models (LMs) rely on their parametric knowledge augmented with relevant contextual knowledge for certain tasks, such as question answering. However, the contextual knowledge can contain private information that may be leaked when answering queries, and estimating this privacy leakage is not well understood. A straightforward approach of directly comparing an LM's output to the contexts can overestimate the privacy risk, since the LM's parametric knowledge might already contain the augmented contextual knowledge. To this end, we introduce \emph{context influence}, a metric that builds on differential privacy, a widely-adopted privacy notion, to estimate the privacy leakage of contextual knowledge during decoding. Our approach effectively measures how each subset of the context influences an LM's response while separating the specific parametric knowledge of the LM. Using our context influence metric, we demonstrate that context privacy leakage occurs when contextual knowledge is out of distribution with respect to parametric knowledge. Moreover, we experimentally demonstrate how context influence properly attributes the privacy leakage to augmented contexts, and we evaluate how factors-- such as model size, context size, generation position, etc.-- affect context privacy leakage. The practical implications of our results will inform practitioners of the privacy risk associated with augmented contextual knowledge.

Watching the Watchers: Exposing Gender Disparities in Machine Translation Quality Estimation

Emmanouil Zaranis, Giuseppe Attanasio, Sweta Agrawal, Andre Martins

Quality estimation (QE)---the automatic assessment of translation quality---has recently become crucial across several stages of the translation pipeline, from data curation to training and decoding. While QE metrics have been optimized to align with human judgments, whether they encode social biases has been largely overlooked. Biased QE risks favoring certain demographic groups over others, e.g., by exacerbating gaps in visibility and usability. This paper defines and investigates gender bias of QE metrics and discusses its downstream implications for machine translation (MT). Experiments with state-of-the-art QE metrics across multiple domains, datasets, and languages reveal significant bias. When a human entity's gender in the source is undisclosed, masculine-inflected translations score higher than feminine-inflected ones, and gender-neutral translations are penalized. Even when contextual cues disambiguate gender, using context-aware QE metrics leads to more errors in selecting the correct translation inflection for feminine referents than for masculine ones. Moreover, a biased QE metric affects data filtering and quality-aware decoding. Our findings underscore the need for a renewed focus on developing and evaluating QE metrics centered on gender.

All That Glitters is Not Novel: Plagiarism in AI Generated Research

Tarun Gupta, Danish Pruthi

Automating scientific research is considered the final frontier of science. Recently, several papers claim autonomous research agents can generate novel research ideas. Amidst the prevailing optimism, we document a critical concern: a considerable fraction of such research documents are smartly plagiarized. Unlike past efforts where experts evaluate the novelty and feasibility of research ideas, we request 13 experts to operate under a different situational logic: to identify similarities between LLM-generated research documents and existing work. Concerningly, the experts identify 24% of the 50 evaluated research documents to be either paraphrased (with one-to-one methodological mapping), or significantly borrowed from existing work. These reported instances are cross-verified by authors of the source papers. Experts find an additional 32% ideas to partially overlap with prior work, and a small fraction to be completely original. Problematically, these LLM-generated research documents do not acknowledge original sources, and bypass inbuilt plagiarism detectors. Lastly, through controlled experiments we show that automated plagiarism detectors are inadequate at catching plagiarized ideas from such systems. We recommend a careful assessment of LLM-generated research, and discuss the implications of our findings on academic publishing.

MTSA: Multi-turn Safety Alignment for LLMs through Multi-round Red-teaming

Weiyang Guo, Jing Li, Wenya Wang, YU LI, Daojing He, Jun Yu, Min Zhang

The proliferation of jailbreak attacks against large language models (LLMs) highlights the need for robust security measures. However, in multi-round dialogues, malicious intentions may be hidden in interactions, leading LLMs to be more prone to produce harmful responses. In this paper, we propose the Multi-Turn Safety Alignment (MTSA) framework, to address the challenge of securing LLMs in multi-round interactions. It consists of two stages: In the thought-guided attack learning stage, the red-team model learns about thought-guided multi-round jailbreak attacks to generate adversarial prompts. In the adversarial iterative optimization stage, the red-team model and the target model continuously improve their respective capabilities in interaction. Furthermore, we introduce a multi-turn reinforcement learning algorithm based on future rewards to enhance the robustness of safety alignment. Experimental results show that the red-team model exhibits state-of-the-art attack capabilities, while the target model significantly improves its performance on safety benchmarks.

From Lists to Emojis: How Format Bias Affects Model Alignment

Xuanchang Zhang, Wei Xiong, Lichang Chen, Tianyi Zhou, Heng Huang, Tong Zhang

In this paper, we study format biases in reinforcement learning from human feedback (RLHF). We observe that many widely-used preference models—including human evaluators, GPT-4, and top-ranking models on the RewardBench benchmark—exhibit strong biases towards specific format patterns, such as lists, links, bold text, and emojis. Furthermore, large language models (LLMs) can exploit these biases to achieve higher rankings on popular benchmarks like AlpacaEval and LMSYS Chatbot Arena. One notable example is verbosity bias, where current preference models favor longer responses that appear more comprehensive, even when their quality is equal to or lower than shorter responses. However, format biases beyond verbosity remain largely underexplored. In this work, we extend the study of biases in preference learning beyond the commonly recognized length bias, offering a comprehensive analysis of a wider range of format biases. Additionally, we show that with a small amount of biased data (less than 1%), we can inject significant bias into the reward model. Moreover, these format biases can also be easily exploited by downstream alignment algorithms, such as *best-of-n sampling* and online iterative *DPO*, as it is usually easier to manipulate the format than to improve the quality of responses. Our findings emphasize the need to disentangle format and content both for designing alignment algorithms and evaluating models.

Different Speech Translation Models Encode and Translate Speaker Gender Differently

Dennis Fucci, Marco Gaido, Matteo Negri, Luisa Bentivogli, Andre Martins, Giuseppe Attanasio

Recent studies on interpreting the hidden states of speech models have shown their ability to capture speaker-specific features, including gender. Does this finding also hold for speech translation (ST) models? If so, what are the implications for the speaker's gender assignment in translation? We address these questions from an interpretability perspective, using probing methods to assess gender encoding across diverse ST models. Results on three language directions (English \rightarrow French/Italian/Spanish) indicate that while traditional encoder-decoder models capture gender information, newer architectures---integrating a speech encoder with a machine translation system via adapters---do not. We also demonstrate that low gender encoding capabilities result in systems' tendency toward a masculine default, a translation bias that is more pronounced in newer architectures.

Data Caricatures: On the Representation of African American Language in Pretraining Corpora

Nicholas Deas, Blake Vente, Amith Ananthram, Jessica A Grieser, Desmond U. Patton, Shana Kleiner, James R. Shepard III, Kathleen McKeown

With a combination of quantitative experiments, human judgments, and qualitative analyses, we evaluate the quantity and quality of African American Language (AAL) representation in 12 predominantly English, open-source pretraining corpora. We specifically focus on the sources, variation, and naturalness of included AAL texts representing the AAL speaking community. We find that AAL is underrepresented in all evaluated pretraining corpora compared to US demographics, constituting as few as 0.007% and at most 0.18% of documents. We also find that more than 25% of AAL texts in C4 may be perceived as inappropriate for LLMs to generate and to reinforce harmful stereotypes. Finally, we find that most automated filters are more likely to conserve White Mainstream English (WME) texts over AAL in pretraining corpora.

R-Fairness: Assessing Fairness of Ranking in Subjective Data

Lorenzo Balzotti, Donatella Firmani, Jerin George Mathew, Riccardo Torlone, Sihem Amer-Yahia

Subjective data, reflecting individual opinions, permeates platforms like Yelp and Amazon, influencing everyday decisions. Upon a user query, collaborative rating platforms return a collection of items ranked in an order that is often not transparent to the users. Then, each item is presented with a collection of reviews in an order that typically is, again, rather opaque. Despite the prevalence of such platforms, little attention has been given to fairness in their context, where groups writing best-ranked reviews for best-ranked items have more influence on users' behavior. We design and evaluate a fairness assessment pipeline that starts with a data collection phase to gather reviews from real-world platforms, by submitting artificial user queries and iterating through rated items. Following that, a group assignment phase computes and infers relevant groups for each review, based on review content and user data. Finally, the third step assesses and evaluates the fairness of rankings for different user groups. The key contributions are comparing group exposure for different queries and platforms and comparing how popular fairness definitions behave in different settings. Experiments on real datasets reveal insights into the impact of item ranking on fairness computation and the varying robustness of these measures.

Session 10: Generation (Tuesday 29 July, 16:00-17:30, Hall 4/5)

Boost, Disentangle, and Customize: A Robust System2-to-System1 Pipeline for Code Generation

Kounianhua Du, Hanjing Wang, Jianxing Liu, Jizheng Chen, Xinyi Dai, Yasheng Wang, Ruiming Tang, Yong Yu, Jun Wang, Weinan Zhang

Large language models (LLMs) have demonstrated remarkable capabilities in various domains, particularly in system 1 tasks, yet the intricacies of their problem-solving mechanisms in system 2 tasks are not sufficiently explored. Recent research on System2-to-System1 methods surge, exploring the System 2 reasoning knowledge via inference-time computation and compressing the explored knowledge into System 1 process. In this paper, we focus on code generation, which is a representative System 2 task, and identify two primary challenges: (1) the complex hidden reasoning processes and (2) the heterogeneous data distributions that complicate the exploration and training of robust LLM solvers. To tackle these issues, we propose a novel BDC framework that explores insightful System 2 knowledge of LLMs using a MC-Tree-Of-Agents algorithm with mutual \textbf{B}oosting, \textbf{D}isentangles the heterogeneous training data for composable LoRA-experts, and obtain \textbf{C}ustomized problem solver for each data instance with an input-aware hypernetwork to weight over the LoRA-experts, offering effectiveness, flexibility, and robustness. This framework leverages multiple LLMs through mutual verification and boosting, integrated into a Monte-Carlo Tree Search process enhanced by reflection-based pruning and refinement. Additionally, we introduce the DisenLora algorithm, which clusters heterogeneous data to fine-tune LLMs into composable Lora experts, enabling the adaptive generation of customized problem solvers through an input-aware hypernetwork. Our contributions include the identification of critical challenges in existing methodologies, the development of the MC-Tree-of-Agents algorithm for insightful data collection, and the creation of a robust and flexible solution for code generation. This work lays the groundwork for advancing LLM capabilities in complex reasoning tasks, offering a novel System2-to-System1 solution.

RAPID: Efficient Retrieval-Augmented Long Text Generation with Writing Planning and Information Discovery

Hongchao Gu, Dexun Li, Kuicai Dong, Hao Zhang, Hang Lv, Hao Wang, Defu Lian, Yong Liu, Enhong Chen

Generating knowledge-intensive and comprehensive long texts, such as encyclopedia articles, remains significant challenges for Large Language Models. It requires not only the precise integration of facts but also the maintenance of thematic coherence throughout the article. Existing methods, such as multi-agent discussion, often struggle with issues like hallucinations, topic incoherence, and significant latency. To address these challenges, we propose RAPID, an efficient **R**etrieval-**A**ugmented long text generation framework with writing **P**lanning and **I**nformation **D**iscovery. RAPID consists of three main modules: (1) Retrieval-augmented preliminary outline generation to reduce hallucinations, (2) Attribute-constrained search for efficient information discovery, (3) Plan-guided article generation for enhanced coherence. Extensive experiments on our newly compiled benchmark dataset, FreshWiki-2024, demonstrate that RAPID significantly outperforms state-of-the-art methods across a wide range of evaluation metrics (\eg long-text generation, outline quality, latency, etc). Our work provides a robust and efficient solution to the challenges of automated long-text generation.

Speculative Decoding via Exponential Races

Szymon Kobus, Deniz Gunduz

Speculative decoding accelerates large language model inference using a smaller draft model. In this paper, we establish a surprising connection between speculative decoding and channel simulation, which aims at simulating a noisy channel using as few bits as possible. This connection allows us to provide an information-theoretic analysis of the speed up that can be achieved by speculative decoding. Leveraging this link, we derive an explicit relation between generation speed-up and the number of tokens k generated by the draft model for large k, which serves as an upper bound for all k. We also propose a novel speculative decoding method via exponential races ERSD that matches state-of-the-art performance.

MetaSynth: Meta-Prompting-Driven Agentic Scaffolds for Diverse Synthetic Data Generation

Haris Riaz, Sourav Sanjukta Bhabesh, Vinayak Arannil, Miguel Ballesteros, Graham Horwood

Recent smaller language models such Phi-3.5 and Phi-4 rely on synthetic data generated using larger language models. Questions remain about leveraging synthetic data for other use cases, such as adapting LLMs to specific domains. A key limitation of synthetic data is low diversity, which negatively impacts its downstream applicability for improving other models. To address this, we propose MetaSynth, a method for generating synthetic data that enhances diversity through meta-prompting, where a language model orchestrates multiple expert LLM agents to collaboratively generate data. Using only 25 million tokens of synthetic data generated with MetaSynth, we successfully adapt a well-trained LLM (Mistral-7B) to two specialized domains--Finance and Biomedicine--without compromising the capabilities of the resulting model in general tasks. In addition, we evaluate the diversity of our synthetic data using seven automated metrics, and find that it approaches the diversity of LLM pre-training corpora.

Confidence Improves Self-Consistency in LLMs

Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, Gal Yona

Self-consistency decoding enhances LLMs' performance on reasoning tasks by sampling diverse reasoning paths and selecting the most frequent answer. However, it is computationally expensive, as sampling many of these (lengthy) paths is required to increase the chances that the correct answer emerges as the most frequent one. To address this, we introduce Confidence-Informed Self-Consistency (CISC). CISC performs a weighted majority vote based on confidence scores obtained directly from the model. By prioritizing high-confidence paths, it can identify the correct answer with a significantly smaller sample size. When tested on nine models and four datasets, CISC outperforms self-consistency in nearly all configurations, reducing the required number of reasoning paths by over 40% on average. In addition, we introduce the notion of within-question confidence evaluation, after showing that standard evaluation methods are poor predictors of success in distinguishing correct and incorrect answers to the same question. In fact, the most calibrated confidence method proved to be the least effective for CISC. Lastly, beyond these practical implications, our results and analyses show that LLMs can effectively judge the correctness of their own outputs, contributing to the ongoing debate on this topic.

Towards A “Novel” Benchmark: Evaluating Literary Fiction with Large Language Models

Wenqing Wang, Mingqi Gao, Xinyu Hu, Xiaojun Wan

Current exploration on creative generation focuses mainly on short stories, poetry, and scripts. With the expansion of Large Language Models (LLMs) context windows, "novel" avenues emerge. This study aims to extend the boundaries of Natural Language Generation (NLG) evaluation by exploring LLMs' capabilities in more challenging long-form fiction. We propose a new multi-level evaluation framework that incorporates ten metrics across the Macro, Meso, and Micro levels. An annotated fiction dataset, sourced from human authors, LLMs, and human-AI collaborations in both English and Chinese is then constructed. Human evaluation reveals notable disparities between LLM-generated and human-authored fictions, particularly the “high-starting, low-ending” pattern in LLM outputs. We further probe ten high-performing LLMs through different prompt templates, achieving moderate correlations by strategically utilizing diverse LLMs tailored to different levels, as an initial step towards better automatic fiction evaluation. Finally, we offer a fine-grained analysis of LLMs capabilities through six issues, providing promising insights for future advancements.

Dolphin: Document Image Parsing via Heterogeneous Anchor Prompting

Hao Feng, Shu Wei, Xiang Fei, Wei Shi, Yingdong Han, Lei Liao, Jinghui Lu, Binghong Wu, Qi Liu, Chunhui Lin, Jingqun Tang, Hao Liu, Can Huang

Document image parsing is challenging due to its complexly intertwined elements such as text paragraphs, figures, formulas, and tables. Current approaches either assemble specialized expert models or directly generate page-level content autoregressively, facing integration overhead, efficiency bottlenecks, and layout structure degradation despite their decent performance. To address these limitations, we present \textit{Dolphin} (\textit{\textbf{Do}cument Image \textbf{P}arsing via \textbf{H}eterogeneous Anchor Prompt\textbf{in}g}), a novel multimodal document image parsing model following an analyze-then-parse paradigm. In the first stage, Dolphin generates a sequence of layout elements in reading order. These heterogeneous elements, serving as anchors and coupled with task-specific prompts, are fed back to Dolphin for parallel content parsing in the second stage. To train Dolphin, we construct a large-scale dataset of over 30 million samples, covering multi-granularity parsing tasks. Through comprehensive evaluations on both prevalent benchmarks and self-constructed ones, Dolphin achieves state-of-the-art performance across diverse page-level and element-level settings, while ensuring superior efficiency through its lightweight architecture and parallel parsing mechanism. The code and pre-trained models are publicly available at \href{https://github.com/ByteDance/Dolphin}{{\tt https://github.com/ByteDance/Dolphin}}

FastDraft: How to Train Your Draft

Ofir Zafrir, Igor Margulis, Dorin Shteyman, Shira Guskin, Guy Boudoukh

Speculative Decoding has gained popularity as an effective technique for accelerating the auto-regressive inference process of Large Language Models. However, Speculative Decoding entirely relies on the availability of efficient draft models, which are often lacking for many existing language models due to a stringent constraint of vocabulary compatibility. In this work we introduce FastDraft, a novel and efficient approach for pre-training and aligning a draft model to any large language model by incorporating efficient pre-training, followed by fine-tuning over synthetic datasets generated by the target model. We demonstrate FastDraft by training two highly parameter efficient drafts for the popular Phi-3-mini and Llama-3.1-8B models. Using FastDraft, we were able to produce a draft with approximately 10 billion tokens on a single server with 8 Intel Gaudi 2 accelerators in under 24 hours. Our results show that the draft model achieves impressive results in key metrics of acceptance rate, block efficiency and up to 3x memory bound speed up when evaluated on code completion and up to 2x in summarization, text completion and instruction tasks. We validate our theoretical findings through benchmarking on the latest Intel Core Ultra, achieving a wall-clock time speedup of up to 2x, indicating a significant reduction in runtime. Due to its high quality, FastDraft unlocks large language models inference on AI-PC and other edge-devices.

Behavioral Analysis of Information Salience in Large Language Models

Jan Trienes, Jörg Schlötterer, Junyi Jessy Li, Christin Seifert

Large Language Models (LLMs) excel at text summarization, a task that requires models to select content based on its importance. However, the exact notion of salience that LLMs have internalized remains unclear. To bridge this gap, we introduce an explainable framework to systematically derive and investigate information salience in LLMs through their summarization behavior. Using length-controlled summarization as a behavioral probe into the content selection process, and tracing the answerability of Questions Under Discussion throughout, we derive a proxy for how models prioritize information. Our experiments on 13 models across four datasets reveal that LLMs have a nuanced, hierarchical notion of salience, generally consistent across model families and sizes. While models show highly consistent behavior and hence salience patterns, this notion of salience cannot be accessed through introspection, and only weakly correlates with human perceptions of information salience.

DIESEL - Dynamic Inference-Guidance via Evasion of Semantic Embeddings in LLMs

Ben Ganon, Alon Zolfi, Omer Hofman, Inderjeet Singh, Hisashi Kojima, Yuval Elovici, Asaf Shabtai

In recent years, large language models (LLMs) have had great success in tasks such as casual conversation, contributing to significant advancements in domains like virtual assistance. However, they often generate responses that are not aligned with human values (e.g., ethical standards, safety), leading to potentially unsafe or inappropriate outputs. While several techniques have been proposed to address this problem, they come with a cost, requiring computationally expensive training or dramatically increasing the inference time. In this paper, we present DIESEL, a lightweight inference-guidance technique that can be seamlessly integrated into any autoregressive LLM to semantically filter undesired concepts from the response. DIESEL can function either as a standalone safeguard or as an additional layer of defense, enhancing response safety by reranking the LLM's proposed tokens based on their similarity to predefined negative concepts in the latent space. Our evaluation demonstrates DIESEL's effectiveness on state-of-the-art conversational models, even in adversarial jailbreaking scenarios that challenge response safety. We also highlight DIESEL's generalization capabilities, showing that it can be used in use cases other than safety, providing general-purpose response filtering.

Rectifying Belief Space via Unlearning to Harness LLMs' Reasoning

Ayana Niwa, Masahiro Kaneko, Kentaro Inui

Large Language Models (LLMs) exhibit sophisticated reasoning yet still generate incorrect answers. We attribute these errors to **Spurious Beliefs**, defined as propositions the model internally considers as true despite being factually false. To reduce reasoning errors, we propose a belief space rectification framework. Our method first identifies the beliefs invoked during inference via an explanation-based approach with Forward-Backward Beam Search (FBBS). We subsequently apply unlearning via gradient ascent to suppress spurious beliefs and enhance true ones, thereby effectively rectifying the model's belief space. Experiments on three QA datasets and three LLMs show that our method significantly reduces erroneous reasoning and improves generalization.

Revisiting Self-Consistency from Dynamic Distributional Alignment Perspective on Answer Aggregation

Yiwei Li, Ji Zhang, Shaoxiong Feng, Peiwen Yuan, Xinglin Wang, Jiayi Shi, Yueqi Zhang, Chuyi Tan, Boyuan Pan, Yao Hu, Kan Li

Self-consistency improves reasoning by aggregating diverse stochastic samples, yet the dynamics behind its efficacy remain underexplored. We reframe self-consistency as a dynamic distributional alignment problem, revealing that decoding temperature not only governs sampling randomness but also actively shapes the latent answer distribution. Given that high temperatures require prohibitively large sample sizes to stabilize, while low temperatures risk amplifying biases, we propose a confidence-driven mechanism that dynamically calibrates temperature: sharpening the sampling distribution under uncertainty to align with high-probability modes, and promoting exploration when confidence is high. Experiments on mathematical reasoning tasks show this approach outperforms fixed-diversity baselines under limited samples, improving both average and best-case performance across varying initial temperatures without additional data or modules. This establishes self-consistency as a synchronization challenge between sampling dynamics and evolving answer distributions.

TagGen: Enforcing Syntactic Structures with Tag-Based Control

Vicky Xefteri, Afra Amini, Tim Vieira, Ryan Cotterell

Instruction fine-tuning has made it easier to control language model outputs for tasks like generating text with a specific topic or sentiment. By using targeted prompts with instructions and examples, instruction-tuned models can generally meet these control objectives. However, controlling the syntactic structure of the generated text remains a challenge. To address this issue, we introduce TagGen, a novel controlled decoding algorithm based on the parsing-as-tagging framework. TagGen uses syntactic taggers to guide the generation process, ensuring that each token aligns with the desired syntactic structure as it is generated. Our experiments with GPT2 and Llama3 models show that TagGen significantly improves syntactic accuracy, increasing the F1 score from 8.34 to 34.63 for GPT2-large and from 36.86 to 47.29 for Llama3. These results highlight the complexity of syntactic control and the effectiveness of TagGen in controlling the syntactic structure of the text, offering a promising approach for applications where precise control over syntax is essential.

PLAY2PROMPT: Zero-shot Tool Instruction Optimization for LLM Agents via Tool Play

Wei Fang, Yang Zhang, Kaizhi Qian, James R. Glass, Yada Zhu

Large language models (LLMs) are increasingly integrated with specialized external tools, yet many tasks demand zero-shot tool usage with minimal or noisy documentation. Existing solutions rely on manual rewriting or labeled data for validation, making them inapplicable in true zero-shot settings. To address these challenges, we propose PLAY2PROMPT, an automated framework that systematically "plays" with each tool to explore its input-output behaviors. Through this iterative trial-and-error process, PLAY2PROMPT refines tool documentation and generates usage examples without any labeled data. These examples not only guide LLM inference but also serve as validation to further enhance tool utilization. Extensive experiments on real-world tasks demonstrate that PLAY2PROMPT significantly improves zero-shot tool performance across both open and closed models, offering a scalable and effective solution for domain-specific tool integration.

A Constrained Text Revision Agent via Iterative Planning and Searching

Hannan Cao, Hwee Tou Ng

Existing text revision systems are capable of generating fluent and coherent text, but struggle with constrained text revision (CTR), which requires adherence to specific constraints. Furthermore, adapting these systems to diverse constraints is challenging. To bridge this gap, we introduce TRIPS, a Text Revision agent via Iterative Planning and Searching, focusing on CTR. TRIPS utilizes a planner, a reviser (i.e., a large language model), and adaptable tools to generate revisions tailored to different scenarios. Specifically, we propose an iterative self-training alignment method to construct the planner, which generates tool usage and text revision plans. Furthermore, we propose Tool-Guided Monte Carlo Tree Search (TG-MCTS), a novel CTR algorithm that extends MCTS with tool-guided expansion and evaluation, enabling the search for optimal revision strategies across various scenarios. To evaluate TRIPS, we introduce ConsTRev, a dataset with multi-level constrained instructions for paragraph-level revision. Experimental results show that TRIPS outperforms baselines in both constraint adherence and revision quality. Furthermore, TRIPS exhibits robust performance across diverse use cases, including plain text and LaTeX revision.

Rationalize and Align: Enhancing Writing Assistance with Rationale via Self-Training for Improved Alignment

Hannan Cao, Hai Ye, Hwee Tou Ng

A Writing Assistant (WA) is a system that offers writing suggestions based on user instructions. Existing WAs are typically built by training large language models (LLMs) on domain-specific instruction data through supervised fine-tuning (SFT) only. However, SFT optimizes models to match a single reference, failing to capture the inherent flexibility of text editing, where multiple valid revisions exist. Therefore, solely relying on SFT limits WA performance. To address this limitation, we propose the Rationalize and Align framework, which enhances the WA performance with rationale (i.e., linguistic explanations) and alignment. Our framework automatically generates the rationale and preference data for writing tasks via distillation and self-training, eliminating the need for human annotation. These data are then leveraged to refine WA using a novel preference optimization method. Empirical results show that our framework significantly improves WA performance. Our WA outperforms both open-source state-of-the-art WAs and the closed-source GPT-4o by 3.9 and 7.1 points on average, respectively, across eight well-established writing-related test sets.

BIPro: Zero-shot Chinese Poem Generation via Block Inverse Prompting Constrained Generation Framework

Xu Zou

Recently, generative pre-trained models have made significant strides, particularly highlighted by the release of ChatGPT and GPT-4, which exhibit superior cross-domain capabilities. However, these models still face challenges on constrained writing tasks like poem generation under open-domain titles via direct generation. In response to this challenge, we introduce Block Inverse Prompting (BIPro) constrained generation framework. BIPro leverages two block inverse prompting methods, revise and rewrite. This inference scaling approach mimics the process of human text writing using block generative models. It significantly improves the zero-shot generation quality on the constrained generation task of open-domain traditional-form Chinese poem generation. Based on a less powerful block generative model GLM-10B-Chinese, poems composed via BIPro without priming or additional training outperform both much larger direct generative systems like GPT-4 or GLM-4 and domain-specific systems such as Yusheng, Shisanbai, or Baidu Poetry Helper in human evaluation by proficient poets. BIPro considerably narrows the gap between AI-generated works and short-listed human literary arts in another human evaluation, unveiling the promising potential of inference scaling in improving the quality of constrained generation. It is open-sourced and available as an agent in chatglm app.

Binary Classifier Optimization for Large Language Model Alignment

Seungjae Jung

In real-world services such as ChatGPT, aligning models based on user feedback is crucial for improving model performance. However, due to the simplicity and convenience of providing feedback, users typically offer only basic binary signals, such as 'thumbs-up' or 'thumbs-down'. Most existing alignment research, on the other hand, relies on preference-based approaches that require both positive and negative responses as a pair. We propose Binary Classifier Optimization (BCO), a technique that effectively aligns LLMs using only binary feedback. BCO trains a binary classifier, where the logit serves as an implicit reward, effectively minimizing the Direct Preference Optimization (DPO) loss. We demonstrate that the binary cross-entropy loss employed in classifier training acts as an upper bound for the DPO loss. Additionally, a novel reward shift technique further minimizes the gap between the losses. We validate our methodology in two settings: first, on a paired preference dataset, where our method performs on par with DPO; and second, on a Likert-5 scale annotation dataset which stems from real users' queries. Our model consistently demonstrates effective and robust alignment across four base LLMs and three different datasets, showcasing the strength of our approach to learning from binary signals.

DioR: Adaptive Cognitive Detection and Contextual Retrieval Optimization for Dynamic Retrieval-Augmented Generation

Hanghui Guo, Jia Zhu, Shimin Di, Weijie Shi, Zhangze Chen, Jiajie Xu

Dynamic Retrieval-augmented Generation (RAG) has shown great success in mitigating hallucinations in large language models (LLMs) during generation. However, existing dynamic RAG methods face significant limitations in two key aspects: 1) Lack of an effective mechanism to control retrieval triggers, and 2) Lack of effective scrutiny of retrieval content. To address these limitations, we propose an innovative dynamic RAG method, DioR (Adaptive Cognitive Detection and Contextual Retrieval Optimization), which consists of two main components: adaptive cognitive detection and contextual retrieval optimization, specifically designed to determine when retrieval is needed and what to retrieve for LLMs is useful. Experimental results demonstrate that DioR achieves superior performance on all tasks, demonstrating the effectiveness of our work.

Revisiting the Test-Time Scaling of o1-like Models: Do they Truly Possess Test-Time Scaling Capabilities?

Zhiyuan Zeng, Qinyuan Cheng, Zhangyue Yin, Yunhua Zhou, Xipeng Qiu

The advent of test-time scaling in large language models (LLMs), exemplified by OpenAI's o1 series, has advanced reasoning capabilities by scaling computational resource allocation during inference. While successors like QwQ, Deepseek-R1 (R1) and LIMO replicate these advancements, whether these models truly possess test-time scaling capabilities remains underexplored. This study found that longer CoTs of these o1-like models do not consistently enhance accuracy; in fact, correct solutions are often shorter than incorrect ones for the same questions. Further investigation shows this phenomenon is closely related to models' self-revision capabilities - longer CoTs contain more self-revisions, which often lead to performance degradation. We then compare sequential and parallel scaling strategies on QwQ, R1 and LIMO, finding that parallel scaling achieves better coverage and scalability. Based on these insights, we propose ``Shortest Majority Vote'', a method that combines parallel scaling strategies with CoT length characteristics, significantly improving models' test-time scalability compared to conventional majority voting approaches.

RPO: Retrieval Preference Optimization for Robust Retrieval-Augmented Generation

Shi-Qi Yan, Quan Liu, Zhen-Hua Ling

While Retrieval-Augmented Generation (RAG) has exhibited promise in utilizing external knowledge, its generation process heavily depends on the quality and accuracy of the retrieved context. Large language models (LLMs) struggle to evaluate the correctness of non-parametric knowledge retrieved externally when it differs from internal memorization, leading to knowledge conflicts during response generation. To this end, we introduce the **R**etrieval **P**reference **O**ptimization (RPO), a lightweight and effective alignment method to adaptively leverage multi-source knowledge based on retrieval relevance. An implicit representation of retrieval relevance is derived and incorporated into the reward model to integrate retrieval evaluation and response generation into a single model, solving the problem that previous methods necessitate the additional procedure to assess the retrieval quality. Notably, RPO is a RAG-dedicated alignment approach that quantifies the awareness of retrieval relevance in training, first overcoming mathematical obstacles. Experiments on four datasets demonstrate that RPO outperforms RAG by 4-10% in accuracy without any extra component, exhibiting its robust generalization.

DTCRS: Dynamic Tree Construction for Recursive Summarization

Guanran Luo

Retrieval-Augmented Generation (RAG) mitigates the hallucination problem of Large Language Models (LLMs) by incorporating external knowledge. Recursive summarization constructs a hierarchical summary tree by clustering text chunks, integrating information from multiple parts of a document to provide evidence for abstractive questions involving multi-step reasoning. However, summary trees often contain a large number of redundant summary nodes, which not only increase construction time but may also negatively impact question answering. Moreover, recursive summarization is not suitable for all types of questions. We introduce DTCRS, a method that dynamically generates summary trees based on document structure and query semantics. DTCRS determines whether a summary tree is necessary by analyzing the question type. It then decomposes the question and uses the embeddings of sub-questions as initial cluster centers, reducing redundant summaries while improving the relevance between summaries and the question. Our approach significantly reduces summary tree construction time and achieves substantial improvements across three QA tasks. Additionally, we investigate the applicability of recursive summarization to different question types, providing valuable insights for future research.

Frictional Agent Alignment Framework: Slow Down and Don’t Break Things

Abhijnan Nath, Carine Graff, Andrei Bachinin, Nikhil Krishnaswamy

AI support of collaborative interactions entails mediating potential misalignment between interlocutor beliefs. Common preference alignment methods like DPO excel in static settings, but struggle in dynamic collaborative tasks where the explicit signals of interlocutor beliefs are sparse and skewed. We propose the Frictional Agent Alignment Framework (FAAF), to generate precise, context-aware "friction" that prompts for deliberation and re-examination of existing evidence. FAAF’s two-player objective decouples from data skew: a frictive-state policy identifies belief misalignments, while an intervention policy crafts collaborator-preferred responses. We derive an analytical solution to this objective, enabling training a single policy via a simple supervised loss. Experiments on three benchmarks show FAAF outperforms competitors in producing concise, interpretable friction and in OOD generalization. By aligning LLMs to act as adaptive "thought partners"---not passive responders---FAAF advances scalable, dynamic human-AI collaboration. Our code and data can be found at https://github.com/csu-signal/FAAF_ACL.

Improving Factuality with Explicit Working Memory

Mingda Chen, Yang Li, Karthik Padthe, Rulin Shao, Alicia Yi Sun, Luke Zettlemoyer, Gargi Ghosh, Wen-tau Yih

Large language models can generate factually inaccurate content, a problem known as hallucination. Recent works have built upon retrieved-augmented generation to improve factuality through iterative prompting but these methods are limited by the traditional RAG design. To address these challenges, we introduce Ewe (Explicit Working Memory), a novel approach that enhances factuality in long-form text generation by integrating a working memory that receives real-time feedback from external resources. The memory is refreshed based on online fact-checking and retrieval feedback, allowing Ewe to rectify false claims during the generation process and ensure more accurate and reliable outputs. Our experiments demonstrate that Ewe outperforms strong baselines on four fact-seeking long-form generation datasets, increasing the factuality metric, VeriScore, by 2 to 6 points absolute without sacrificing the helpfulness of the responses. Further analysis reveals that the design of rules for memory updates, configurations of memory units, and the quality of the retrieval datastore are crucial factors for influencing model performance.

HyKGE: A Hypothesis Knowledge Graph Enhanced RAG Framework for Accurate and Reliable Medical LLMs Responses

Xinke Jiang, Ruizhe Zhang, Yongxin Xu, Rihong Qiu, Yue Fang, Zhiyuan Wang, Jinyi Tang, Hongxin Ding, Xu Chu, Junfeng Zhao, Yasha Wang

In this paper, we investigate the retrieval-augmented generation (RAG) based on Knowledge Graphs (KGs) to improve the accuracy and reliability of Large Language Models (LLMs). Recent approaches suffer from insufficient and repetitive knowledge retrieval, tedious and time-consuming query parsing, and monotonous knowledge utilization. To this end, we develop a Hypothesis Knowledge Graph Enhanced (HyKGE) framework, which leverages LLMs' powerful reasoning capacity to compensate for the incompleteness of user queries, optimizes the interaction process with LLMs, and provides diverse retrieved knowledge. Specifically, HyKGE explores the zero-shot capability and the rich knowledge of LLMs with Hypothesis Outputs to extend feasible exploration directions in the KGs, as well as the carefully curated prompt to enhance the density and efficiency of LLMs' responses. Furthermore, we introduce the HO Fragment Granularity-aware Rerank Module to filter out noise while ensuring the balance between diversity and relevance in retrieved knowledge. Experiments on two Chinese medical multiple-choice question datasets and one Chinese open-domain medical Q\&A dataset with two LLM turbos demonstrate the superiority of HyKGE in terms of accuracy and explainability. Code is available at https://github.com/Artessay/HyKGE.

SAM Decoding: Speculative Decoding via Suffix Automaton

Yuxuan Hu, Ke Wang, Xiaokang Zhang, Fanjin Zhang, Cuiping Li, Hong Chen, Jing Zhang

Speculative decoding (SD) has been demonstrated as an effective technique for lossless LLM inference acceleration. Retrieval-based SD methods, one kind of model-free method, have yielded promising speedup, but they often rely on single retrieval resources, inefficient retrieval methods, and are constrained to certain tasks. This paper presents a novel retrieval-based speculative decoding method that adapts the suffix automaton (SAM) for efficient and accurate draft generation by utilizing the generating text sequence and static text corpus. Unlike existing n-gram matching methods, SAM-Decoding finds the exact longest suffix match, achieving an average time complexity of O(1) per generation step of SAM update and suffix retrieval. It can also integrate with existing methods, adaptively selecting a draft generation strategy based on match length to generalize to broader domains. Extensive experiments on Spec-Bench show that our method is 18\% faster than other retrieval-based SD methods. Additionally, when combined with advanced EAGLE-2, it provides an additional speedup of 3.28\% -- 11.13\% across various-sized LLM backbones.

GiFT: Gibbs Fine-Tuning for Code Generation

Haochen Li, Wanjin Feng, Xin Zhou, Zhiqi Shen

Training Large Language Models (LLMs) with synthetic data is a prevalent practice in code generation. A key approach is self-training, where LLMs are iteratively trained on self-generated correct code snippets. In this case, the self-generated codes are drawn from a conditional distribution, conditioned on a specific seed description. However, the seed description is not the only valid representation that aligns with its intended meaning. With all valid descriptions and codes forming a joint space, codes drawn from the conditional distribution would lead to an underrepresentation of the full description-code space. As such, we propose Gibbs Fine-Tuning (GiFT), a novel self-training method inspired by Gibbs sampling. GiFT allows self-generated data to be drawn from the marginal distribution of the joint space, thereby mitigating the biases inherent in conditional sampling. We provide a theoretical analysis demonstrating the potential benefits of fine-tuning LLMs with code derived from the marginal distribution. Furthermore, we propose a perplexity-based code selection method to mitigate the imbalanced long-tail distribution of the self-generated codes. Empirical evaluation of two LLMs across four datasets demonstrates that GiFT achieves superior performance, particularly on more challenging benchmarks. Source code is available at .

RankCoT: Refining Knowledge for Retrieval-Augmented Generation through Ranking Chain-of-Thoughts

Mingyan Wu, Zhenghao Liu, Yukun Yan, Xinze Li, Shi Yu, Zheni Zeng, Yu Gu, Ge Yu

Retrieval-Augmented Generation (RAG) enhances the performance of Large Language Models (LLMs) by incorporating external knowledge. However, LLMs still encounter challenges in effectively utilizing the knowledge from retrieved documents, often being misled by irrelevant or noisy information. To address this issue, we introduce RankCoT, a knowledge refinement method that incorporates reranking signals in generating CoT-based summarization for knowledge refinement based on given query and all retrieval documents. During training, RankCoT prompts the LLM to generate Chain-of-Thought (CoT) candidates based on the query and individual documents. It then fine-tunes the LLM to directly reproduce the best CoT from these candidate outputs based on all retrieved documents, which requires LLM to filter out irrelevant documents during generating CoT-style summarization. Additionally, RankCoT incorporates a self-reflection mechanism that further refines the CoT outputs, resulting in higher-quality training data. Our experiments demonstrate the effectiveness of RankCoT, showing its superior performance over other knowledge refinement models. Further analysis reveals that RankCoT can provide shorter but effective refinement results, enabling the generator to produce more accurate answers. All code and data are available at https://github.com/NEUIR/RankCoT.

UniRAG: Unified Query Understanding Method for Retrieval Augmented Generation

Rui Li, Liyang He, Qi Liu, Zheng Zhang, Heng Yu, Yuyang Ye, Linbo Zhu, Yu Su

Retrieval-Augmented Generation (RAG) technology effectively addresses the issues of knowledge update lag and hallucinations in large language models (LLMs) by integrating internal and external knowledge. Existing query augmentation methods improve RAG's performance in handling complex queries but face two key challenges: (1) the separation of query augmentation and encoding tasks, which hinders information sharing and introduces cumulative errors, and (2) the difficulty of selecting the optimal augmentation strategy for different scenarios. In this work, we propose UniRAG, a unified framework for query understanding in RAG. UniRAG employs a decoder-only LLM to jointly perform query augmentation and encoding, eliminating task separation. To facilitate adaptive query augmentation, we categorize existing techniques into query paraphrasing, query expansion, and query abstraction. Our model learns to select the optimal augmentation strategy based on user queries, leveraging retrieval and generation outputs as feedback. Experimental results show that UniRAG significantly outperforms traditional query augmentation methods in five knowledge-intensive benchmark tasks in both closed and open domain question answering.

SCULPT: Systematic Tuning of Long Prompts

Shanu Kumar, Akhila Yesantarao Venkata, Shubhanshu Khandelwal, Bishal Santra, Parag Agrawal, Manish Gupta

Prompt optimization is essential for effective utilization of large language models (LLMs) across diverse tasks. While existing optimization methods are effective in optimizing short prompts, they struggle with longer, more complex ones, often risking information loss and being sensitive to small perturbations. To address these challenges, we propose SCULPT (Systematic Tuning of Long Prompts), a framework that treats prompt optimization as a hierarchical tree refinement problem. SCULPT represents prompts as tree structures, enabling targeted modifications while preserving contextual integrity. It employs a Critic-Actor framework that generates reflections and applies actions to refine the prompt. Evaluations demonstrate SCULPT's effectiveness on long prompts, its robustness to adversarial perturbations, and its ability to generate high-performing prompts even without any initial human-written prompt. Compared to existing state of the art methods, SCULPT consistently improves LLM performance by preserving essential task information while applying structured refinements. Both qualitative and quantitative analyses show that SCULPT produces more stable and interpretable prompt modifications, ensuring better generalization across tasks.

Mutual-Taught for Co-adapting Policy and Reward Models

Tianyuan Shi, Canbin Huang, Fanqi Wan, Longguang Zhong, Ziyi Yang, Weizhou Shen, Xiaojun Quan, Ming Yan

During the preference optimization of large language models (LLMs), distribution shifts may arise between newly generated model samples and the data used to train the reward model (RM). This shift reduces the efficacy of the RM, which in turn negatively impacts the performance of the policy model (PM). To address this challenge, we propose \textbf{Mutual-Taught}, a self-training method that iteratively improves both the PM and RM without requiring additional human annotation. Our approach mirrors the expectation-maximization (EM) algorithm. In the E-step, the PM is updated using feedback from the current RM, guiding the PM toward a better approximation of the latent optimal preference distribution. In the M-step, we update the RM by constructing training data from the outputs of the PM before and after the E-step update. This process ensures that the RM adapts to the evolving policy distribution. Experimental results demonstrate that this iterative approach leads to consistent improvements in both models. Specifically, our 8B policy model, LLaMA-3-8B-Instruct-MT, achieves a length-controlled win rate of 54.1\% on AlpacaEval-2, while our 8B reward model, FsfairX-LLaMA3-RM-MT, performs on par with GPT-4o-2024-08-06 on RewardBench.

Length Controlled Generation for Black-box LLMs

Yuxuan Gu, Wenjie Wang, Xiaocheng Feng, Weihong Zhong, kun Zhu, Lei Huang, Ting Liu, Bing Qin, Tat-Seng Chua

Large language models (LLMs) have demonstrated impressive instruction following capabilities, while still struggling to accurately manage the length of the generated text, which is a fundamental requirement in many real-world applications. Existing length control methods involve fine-tuning the parameters of LLMs, which is inefficient and suboptimal for practical use. In this paper, we propose a novel iterative sampling framework for text length control, integrating the Metropolis-Hastings algorithm with an importance sampling acceleration strategy. This framework efficiently and reliably regulates LLMs to generate length-constrained text without modifying the underlying parameters, thereby preserving the original capabilities of LLMs. Experimental results demonstrate that our framework achieves almost 100% success rates of length control on Llama3.1 for tasks such as length-controlled abstractive summarization and length-constrained instruction following, with minimal additional computational overhead. This also highlights the significant potential of our method for precise length control across a broader range of applications, without compromising the versatility of LLMs.

EventRAG: Enhancing LLM Generation with Event Knowledge Graphs

Zairun Yang, Yilin Wang, Zhengyan Shi, Yuan Yao, Lei Liang, Keyan Ding, Emine Yilmaz, Huajun Chen, Qiang Zhang

Retrieval-augmented generation (RAG) systems often struggle with narrative-rich documents and event-centric reasoning, particularly when synthesizing information across multiple sources. We present EventRAG, a novel framework that enhances text generation through structured event representations. We first construct an Event Knowledge Graph by extracting events and merging semantically equivalent nodes across documents, while expanding under-connected relationships. We then employ an iterative retrieval and inference strategy that explicitly captures temporal dependencies and logical relationships across events. Experiments on UltraDomain and MultiHopRAG benchmarks show EventRAG's superiority over baseline RAG systems, with substantial gains in generation effectiveness, logical consistency, and multi-hop reasoning accuracy. Our work advances RAG systems by integrating structured event semantics with iterative inference, particularly benefiting scenarios requiring temporal and logical reasoning across documents.

Safer or Luckier? LLMs as Safety Evaluators Are Not Robust to Artifacts

Hongyu Chen, Seraphina Goldfarb-Tarrant

Large Language Models (LLMs) are increasingly employed as automated evaluators to assess the safety of generated content, yet their reliability in this role remains uncertain. This study evaluates a diverse set of 11 LLM judge models across critical safety domains, examining three key aspects: self-consistency in repeated judging tasks, alignment with human judgments, and susceptibility to input artifacts such as apologetic or verbose phrasing. Our findings reveal that biases in LLM judges can significantly distort the final verdict on which content source is safer, undermining the validity of comparative evaluations. Notably, apologetic language artifacts alone can skew evaluator preferences by up to 98%. Contrary to expectations, larger models do not consistently exhibit greater robustness, while smaller models sometimes show higher resistance to specific artifacts. To mitigate LLM evaluator robustness issues, we investigate jury-based evaluations aggregating decisions from multiple models. Although this approach both improves robustness and enhances alignment to human judgements, artifact sensitivity persists even with the best jury configurations. These results highlight the urgent need for diversified, artifact-resistant methodologies to ensure reliable safety assessments.

CoreEval: Automatically Building Contamination-Resilient Datasets with Real-World Knowledge toward Reliable LLM Evaluation

Jingqian Zhao, Bingbing Wang, Geng Tu, Yice Zhang, Qianlong Wang, Bin Liang, Jing Li, Ruifeng Xu

Data contamination poses a significant challenge to the fairness of LLM evaluations in natural language processing tasks by inadvertently exposing models to test data during training. Current studies mitigate this issue by modifying existing datasets or generating new ones from freshly collected information. However, these methods fall short of ensuring contamination-resilient evaluation, as they fail to fully eliminate pre-existing knowledge from models or preserve the semantic complexity of the original datasets. To address these limitations, we propose \textbf{CoreEval}, a \textbf{Co}ntamination-\textbf{re}silient \textbf{Eval}uation strategy for automatically updating data with real-world knowledge. This approach begins by extracting entity relationships from the original data and leveraging the GDELT database to retrieve relevant and up-to-date knowledge. The retrieved knowledge is then recontextualized and integrated with the original data, which is refined and restructured to ensure semantic coherence and enhanced task relevance. Ultimately, a robust data reflection mechanism in a Chain-of-Thought manner to iteratively verify and refine labels, ensuring consistency between the updated and original datasets. Extensive experiments on updated datasets validate the robustness of CoreEval, demonstrating its effectiveness in mitigating performance overestimation caused by data contamination.

Uncertainty-Aware Iterative Preference Optimization for Enhanced LLM Reasoning

Lei Li, Hehuan Liu, Yaxin Zhou, ZhaoYang Gui, Xudong Weng, Yi YUAN, Zheng Wei, Zang Li

Direct Preference Optimization (DPO) has recently emerged as an efficient and effective method for aligning large language models with human preferences. However, constructing high-quality preference datasets remains challenging, often necessitating expensive manual or powerful LM annotations. Additionally, standard DPO exhibits suboptimal performance in complex reasoning tasks, such as mathematical and code reasoning. In this paper, we introduce an approach to collect preference pairs through iterative sampling and execution feedback, tailored to the current learning state (\emph{e.g.} well-learned, mis-learned, and unlearned) of the policy model. To alleviate the failures of DPO and improve its applicability in reasoning tasks, we propose {\ours}, an iterative uncertainty-aware preference optimization method that achieves fine-grained preference control by assessing model confidence. We validate our approach across three reasoning tasks, incorporating five established reasoning datasets and one self-curated dataset. Our experimental results demonstrate an overall improvement of 3.6% over the standard DPO method and show the model exhibits promising generalizability.

Personalized Generation In Large Model Era: A Survey

Yiyan Xu, Jinghao Zhang, Alireza Salemi, Xinting Hu, Wenjie Wang, Fuli Feng, Hamed Zamani, Xiangnan He, Tat-Seng Chua

In the era of large models, content generation is gradually shifting to Personalized Generation (PGen), tailoring content to individual preferences and needs. This paper presents the first comprehensive survey on PGen, investigating existing research in this rapidly growing field. We conceptualize PGen from a unified perspective, systematically formalizing its key components, core objectives, and abstract workflows. Based on this unified perspective, we propose a multi-level taxonomy, offering an in-depth review of technical advancements, commonly used datasets, and evaluation metrics across multiple modalities, personalized contexts, and tasks. Moreover, we envision the potential applications of PGen and highlight open challenges and promising directions for future exploration. By bridging PGen research across multiple modalities, this survey serves as a valuable resource for fostering knowledge sharing and interdisciplinary collaboration, ultimately contributing to a more personalized digital landscape.

Help Me Write a Story: Evaluating LLMs' Ability to Generate Writing Feedback

Hannah Rashkin, Elizabeth Clark, Fantine Huot, Mirella Lapata

Can LLMs provide support to creative writers by giving meaningful writing feedback? In this paper, we explore the challenges and limitations of model-generated writing feedback by defining a new task, dataset, and evaluation frameworks. To study model performance in a controlled manner, we present a novel test set of 1,300 stories that we corrupted to intentionally introduce writing issues. We study the performance of commonly used LLMs in this task with both automatic and human evaluation metrics. Our analysis shows that current models have strong out-of-the-box behavior in many respects—providing specific and mostly accurate writing feedback. However, models often fail to identify the biggest writing issue in the story and to correctly decide when to offer critical vs. positive feedback.

Balancing Diversity and Risk in LLM Sampling: How to Select Your Method and Parameter for Open-Ended Text Generation

Yuxuan Zhou, Margret Keuper, Mario Fritz

Sampling-based decoding strategies have been widely adopted for Large Language Models (LLMs) in numerous applications, targeting a balance between diversity and quality via temperature tuning and tail truncation. Considering the strong dependency of the candidate next tokens on different prefixes, recent studies propose to adaptively truncate the tail of LLMs' predicted distribution. Although improved results have been reported with these methods on open-ended text generation tasks, the results are highly dependent on the curated parameters and the limited exemplar text. In this paper, we propose a systematic way to estimate the intrinsic capacity of a truncation sampling method by considering the trade-off between diversity and risk at each decoding step, based on our collected prefix tree which preserves the context of a full sentence. Our work offers a comprehensive comparison of existing truncation sampling methods and serves as a practical user guideline for their parameter selection. Our code is available at https://anonymous.4open.science/r/Truncation-Sampling-Evaluation-251F.

QuASAR: A Question-Driven Structure-Aware Approach for Table-to-Text Generation

WeiJie Liu, Yibin Zheng, Fang Kong

Table-to-text generation aims to automatically produce natural language descriptions from structured or semi-structured tabular data. Unlike traditional text generation tasks, it requires models to accurately understand and represent table structures. Existing approaches typically process tables by linearizing them or converting them into graph structures. However, these methods either fail to adequately capture the table structure or rely on complex attention mechanisms, limiting their applicability. To tackle these challenges, we propose QuASAR, a question-driven self-supervised approach designed to enhance the model's structural perception and representation capabilities. Specifically, QuASAR formulates a set of structure-related queries for self-supervised training, explicitly guiding the model to capture both local and global table structures. Additionally, we introduce two auxiliary pre-training tasks: a word-to-sentence reconstruction task and a numerical summarization task, which further enhance the fluency and factuality of the generated text. Experimental results on the ToTTo and HiTab datasets demonstrate that our approach produces higher-quality text compared to existing methods.

ProgCo: Program Helps Self-Correction of Large Language Models

Xiaoshuai Song, Yanan Wu, Weixun Wang, Jiaheng Liu, Wenbo Su, Bo Zheng

Self-Correction aims to enable large language models (LLMs) to self-verify and self-refine their initial responses without external feedback. However, LLMs often fail to effectively self-verify and generate correct feedback, further misleading refinement and leading to the failure of self-correction, especially in complex reasoning tasks. In this paper, we propose Program-driven Self-Correction (ProgCo). First, program-driven verification (ProgVe) achieves complex verification logic and extensive validation through self-generated, self-executing verification pseudo-programs. Then,program-driven refinement (ProgRe) receives feedback from ProgVe, conducts dual reflection and refinement on both responses and verification programs to mitigate misleading of incorrect feedback in complex reasoning tasks. Experiments on three instruction-following and mathematical benchmarks indicate that ProgCo achieves effective self-correction, and can be further enhance performance when combined with real program tools. We release our code at https://github.com/songxiaoshuai/progco.

Lexical Diversity-aware Relevance Assessment for Retrieval-Augmented Generation

Zhange Zhang, Yuqing Ma, Yulong Wang, Shan He, Tianbo Wang, Siqi He, Jiakai Wang, Xianglong Liu

Retrieval-Augmented Generation (RAG) has proven effective in enhancing the factuality of LLMs' generation, making them a focal point of research. However, previous RAG approaches overlook the lexical diversity of queries, hindering their ability to achieve a granular relevance assessment between queries and retrieved documents, resulting in suboptimal performance. In this paper, we introduce a Lexical Diversity-aware RAG (DRAG) method to address the biases in relevant information retrieval and utilization induced by lexical diversity. Specifically, a Diversity-sensitive Relevance Analyzer is proposed to decouple and assess the relevance of different query components (words, phrases) based on their levels of lexical diversity, ensuring precise and comprehensive document retrieval. Moreover, a Risk-guided Sparse Calibration strategy is further introduced to calibrate the generated tokens that is heavily affected by irrelevant content. Through these modules, DRAG is capable of effectively retrieving relevant documents and leverages their pertinent knowledge to refine the original results and generate meaningful outcomes. Extensive experiments on widely used benchmarks demonstrate the efficacy of our approach, yielding a 10.6% accuracy improvement on HotpotQA.

Subtle Errors in Reasoning: Preference Learning via Error-injected Self-editing

Kaishuai Xu, Tiezheng YU, Wenjun Hou, Yi Cheng, Chak Tou Leong, Liangyou Li, Xin Jiang, Lifeng Shang, Qun Liu, Wenjie Li

Large Language Models (LLMs) have exhibited strong mathematical reasoning prowess, tackling tasks ranging from basic arithmetic to advanced competition-level problems. However, frequently occurring subtle yet critical errors, such as miscalculations or incorrect substitutions, limit the LLMs' full potential. Existing studies to improve mathematical ability typically involve applying preference learning to step-wise solution pairs. Although these methods leverage samples of varying granularity to mitigate reasoning errors, they overlook critical subtle errors. In this work, we propose a novel preference learning framework called eRror-Injected Self-Editing (RISE), which injects predefined subtle errors into pivotal tokens in reasoning or computation steps to construct hard pairs for error mitigation. In detail, RISE uses the LLM itself to edit a small number of tokens in the solution, injecting designed subtle errors. Then, pairs composed of self-edited solutions and their corresponding correct ones, along with pairs of correct and incorrect solutions obtained through sampling, are used together for subtle error-aware DPO training. Compared with other preference learning methods, RISE further refines the training objective without requiring fine-grained sampling or preference annotation. Extensive experiments validate the effectiveness of RISE, with preference learning on Qwen2-7B-Instruct yielding notable improvements of 3.0% on GSM8K and 7.9% on MATH with only 4.5K training samples. Moreover, the effect of error mitigation extends from mathematical reasoning to logical reasoning and code generation.

DualRAG: A Dual-Process Approach to Integrate Reasoning and Retrieval for Multi-Hop Question Answering

Rong Cheng, Jinyi Liu, YAN ZHENG, Fei Ni, Jiazhen Du, Hangyu Mao, Fuzheng Zhang, Bo Wang, Jianye HAO

Multi-Hop Question Answering (MHQA) tasks permeate real-world applications, posing challenges in orchestrating multi-step reasoning across diverse knowledge domains. While existing approaches have been improved with iterative retrieval, they still struggle to identify and organize dynamic knowledge. To address this, we propose DualRAG, a synergistic dual-process framework that seamlessly integrates reasoning and retrieval. DualRAG operates through two tightly coupled processes: Reasoning-augmented Querying (RaQ) and progressive Knowledge Aggregation (pKA). They work in concert: as RaQ navigates the reasoning path and generates targeted queries, pKA ensures that newly acquired knowledge is systematically integrated to support coherent reasoning. This creates a virtuous cycle of knowledge enrichment and reasoning refinement. Through targeted fine-tuning, DualRAG preserves its sophisticated reasoning and retrieval capabilities even in smaller-scale models, demonstrating its versatility and core advantages across different scales. Extensive experiments demonstrate that this dual-process approach substantially improves answer accuracy and coherence, approaching, and in some cases surpassing, the performance achieved with oracle knowledge access. These results establish DualRAG as a robust and efficient solution for complex multi-hop reasoning tasks.

Towards Economical Inference: Enabling DeepSeek's Multi-Head Latent Attention in Any Transformer-based LLMs

Tao Ji, Bin Guo, Yuanbin Wu, Qipeng Guo, shenlixing, chenzhan, Xipeng Qiu, Qi Zhang, Tao Gui

Multi-head Latent Attention (MLA) is an innovative architecture proposed by DeepSeek, designed to ensure efficient and economical inference by significantly compressing the Key-Value (KV) cache into a latent vector. Compared to MLA, standard LLMs employing Multi-Head Attention (MHA) and its variants such as Grouped-Query Attention (GQA) exhibit significant cost disadvantages. Enabling well-trained LLMs (e.g., Llama) to rapidly adapt to MLA without pre-training from scratch is both meaningful and challenging. This paper proposes the first data-efficient fine-tuning method for transitioning from MHA to MLA (**MHA2MLA**), which includes two key components: for *partial-RoPE*, we remove RoPE from dimensions of queries and keys that contribute less to the attention scores, for *low-rank approximation*, we introduce joint SVD approximations based on the pre-trained parameters of keys and values. These carefully designed strategies enable MHA2MLA to recover performance using only a small fraction (0.6\% to 1\%) of the data, significantly reducing inference costs while seamlessly integrating with compression techniques such as KV cache quantization. For example, the KV cache size of Llama2-7B is reduced by 92.19\%, with only a 1\% drop in LongBench performance. Our source code is publicly available at https://github.com/JT-Ushio/MHA2MLA.

Session 10: Generation (Tuesday 29 July, 16:00-17:30, Hall 5X)

ATGen: A Framework for Active Text Generation

Akim Tsvigun, Daniil Vasilev, Ivan Tsvigun, Ivan Lysenko, Talgat Bektleuov, Aleksandr Medvedev, Nikita Severin, Mikhail Mozikov, Andrey Savchenko, Ilya Makarov, Grigorev Rostislav, Ramil Kuleev, Fedor Zhdanov, Artem Shelmanov

Active learning (AL) has demonstrated remarkable potential in reducing the annotation effort required for training machine learning models. However, despite the surging popularity of natural language generation (NLG) tasks in recent years, the application of AL to NLG has been limited. In this paper, we introduce Active Text Generation (ATGen) -- a comprehensive framework that bridges AL with text generation tasks, enabling the application of state-of-the-art AL strategies to NLG. Our framework simplifies AL-empowered annotation in NLG tasks using both human annotators and automatic annotation agents based on large language models (LLMs). The framework supports LLMs deployed as a service, such as ChatGPT and Claude, or operated on-premises. Furthermore, ATGen provides a unified platform for smooth implementation and benchmarking of novel AL strategies tailored to NLG tasks. Finally, we present experimental results on several text generation tasks where we compare the performance of state-of-the-art AL strategies in various settings. We demonstrate that ATGen can reduce both the effort of human annotators and costs for API calls to automatic annotation agents based on LLMs.

Session 10: Industry (Tuesday 29 July, 16:00-17:30, Hall 4X)

Multi-Step Generation of Test Specifications using Large Language Models for System-Level Requirements

Dragan Milchevski, Gordon Frank, Anna Hätty, Bingqing Wang, Xiaowei Zhou, Zhe Feng

System-level testing is a critical phase in the development of large, safety-dependent systems, such as those in the automotive industry. However, creating test specifications can be a time-consuming and error-prone process. This paper presents an AI-based assistant to aid users in creating test specifications for system-level requirements. The system mimics the working process of a test developer by utilizing a LLM and an agentic framework, and by introducing intermediate test artifacts - structured intermediate representations derived from input requirements. Our user study demonstrates a 30 to 40% reduction in effort required for test development. For test specification generation, our quantitative analysis reveals that iteratively providing the model with more targeted information, like examples of similar test specifications, based on comparable requirements and purposes, can boost the performance by up to 30% in ROUGE-L. Overall, our approach has the potential to improve the efficiency, accuracy, and reliability of system-level testing and can be applied to various industries where safety and functionality are paramount.

SocialForge: simulating the social internet to provide realistic training against influence operations

Ulysse Oliveri, Guillaume Gadek, Alexandre Dey, Benjamin Costé, Damien Lolive, Arnaud Delhay, Bruno Grilheres

Social media platforms have enabled large-scale influence campaigns, impacting democratic processes. To fight against these threats, continuous training is needed. A typical training session is based on a fictive scenario describing key elements which are instantiated into a dedicated platform. Such a platform simulates social networks, which host a huge amount of content aligned with the training scenario. However, directly using Large Language Models to create appropriate content result in low content diversity due to coarse-grained and high-level scenario constraints, which compromises the trainees' immersion. We address this issue with SocialForge, a system designed to enhance the diversity and realism of the generated content while ensuring its adherence to the original scenario. Specifically, SocialForge refines and augments the initial scenario constraints by generating detailed subnarratives, personas, and events. We assess diversity, realism, and adherence to the scenario through custom evaluation protocol. We also propose an automatic method to detect erroneous constraint generation, ensuring optimal alignment of the content with the scenario. SocialForge has been used in real trainings and in several showcases, with great end-user satisfaction. We release an open-source dataset generated with SocialForge for the research community.

Genetic Instruct: Scaling up Synthetic Generation of Coding Instructions for Large Language Models

Somshubra Majumdar, Vahid Noroozi, Mehrzad Samadiarakhshbahar, Sean Narenthiran, Aleksander Ficek, Wasi Uddin Ahmad, Jocelyn Huang, Jagadeesh Balam, Boris Ginsburg

Large Language Models (LLMs) require high quality instruction data for effective alignment, particularly in code generation tasks where expert curated datasets are expensive to produce. We present Genetic-Instruct, a scalable algorithm for synthesizing large-scale, high quality coding instructions using evolutionary principles. Starting from a small set of seed instructions, Genetic-Instruct generates diverse and challenging instruction-code pairs by leveraging an Instructor-LLM for generation, a Coder-LLM for code synthesis, and a Judge-LLM for automatic quality evaluation. Our proposed approach is highly parallelizable and effective even with a small seed data and weaker generator models. We generated more than 7.5 million coding instructions with the proposed approach. Then we evaluated it by fine-tuning LLMs with the synthetic samples and demonstrated a significant improvement in their code generation capability compared to the other synthetic generation approaches and publicly available datasets. Our results highlight the efficiency, scalability, and generalizability of the Genetic-Instruct framework.

Generating OpenAPI Specifications from Online API Documentation with Large Language Models

Koren Lazar, Matan Vetzler, Kiran Kate, Jason Tsay, David Boaz, Himanshu Gupta, Avraham Shinnar, Rohith D Vallam, David Amid, Esther Goldbraich, Jim Laredo, Ateret Anaby Tavor

AI agents and business automation tools interacting with external web services require standardized, machine-readable information about their APIs in the form of API specifications. However, the information about APIs available online is often presented as unstructured, free-form HTML documentation, requiring external users to spend significant time manually converting it into a structured format. To address this, we introduce \productname{}, a novel framework that transforms long and diverse API documentation pages into consistent, machine-readable API specifications. This is achieved through a carefully crafted pipeline that integrates large language models and rule-based algorithms which are guided by domain knowledge of the structure of documentation webpages. Our experiments demonstrate that \productname{} generalizes well across hundreds of APIs, and produces valid OpenAPI specifications that encapsulate most of the information from the original documentation. \productname{} has been successfully implemented in an enterprise environment, saving thousands of hours of manual effort and making hundreds of complex enterprise APIs accessible as tools for LLMs.

Graph-Linguistic Fusion: Using Language Models for Wikidata Vandalism Detection

Mykola Trokhymovych, Lydia Pintscher, Ricardo Baeza-Yates, Diego Sáez Trumper

We introduce a next-generation vandalism detection system for Wikidata, one of the largest open-source structured knowledge bases on the Web. Wikidata is highly complex: its items incorporate an ever-expanding universe of factual triples and multilingual texts. While edits can alter both structured and textual content, our approach converts all edits into a single space using a method we call Graph2Text. This allows for evaluating all content changes for potential vandalism using a single multilingual language model. This unified approach improves coverage and simplifies maintenance. Experiments demonstrate that our solution outperforms the current production system. Additionally, we are releasing the code under an open license along with a large dataset of various human-generated knowledge alterations, enabling further research.

LOTUS: A Leaderboard for Detailed Image Captioning from Quality to Societal Bias and User Preferences

Yusuke Hirota, Boyi Li, Ryo Hachiuma, Yueh-Hua Wu, Boris Ivanovic, Marco Pavone, Yejin Choi, Yu-Chiang Frank Wang, Yuta Nakashima, Chao-Han Huck Yang

Large Vision-Language Models (LVLMs) have transformed image captioning, shifting from concise captions to detailed descriptions. We introduce LOTUS, a leaderboard for evaluating detailed captions, addressing three main gaps in existing evaluations: lack of standardized criteria, bias-aware assessments, and user preference considerations. LOTUS comprehensively evaluates various aspects, including caption quality (e.g., alignment, descriptiveness), risks (e.g., hallucination), and societal biases (e.g., gender bias) while enabling preference-oriented evaluations by tailoring criteria to diverse user preferences. Our analysis of recent LVLMs reveals no single model excels across all criteria, while correlations emerge between caption detail and bias risks. Preference-oriented evaluations demonstrate that optimal model selection depends on user priorities.

Efficient Out-of-Scope Detection in Dialogue Systems via Uncertainty-Driven LLM Routing

Alvaro Zaera, Diana Nicoleta Popa, Ivan Sekulic, Paolo Rosso

Out-of-scope (OOS) intent detection is a critical challenge in task-oriented dialogue systems (TODS), as it ensures robustness to unseen and ambiguous queries. In this work, we propose a novel but simple modular framework that combines uncertainty modeling with fine-tuned large language models (LLMs) for efficient and accurate OOS detection. The first step applies uncertainty estimation to the output of an in-scope intent detection classifier, which is currently deployed in a real-world TODS handling tens of thousands of user interactions daily. The second step then leverages an emerging LLM-based approach, where a fine-tuned LLM is triggered to make a final decision on instances with high uncertainty. Unlike prior approaches, our method effectively balances computational efficiency and performance, combining traditional approaches with LLMs and yielding state-of-the-art results on key OOS detection benchmarks, including real-world OOS data acquired from a deployed TODS.

A Perspective on LLM Data Generation with Few-shot Examples: from Intent to Kubernetes Manifest

Antonino Angi, Liubov Nedoshivina, Alessio Sacco, Stefano Braghin, Mark Purcell

The advent of Large Language Models (LLMs) has transformed how complex tasks across various domains can be automated. One of the industry trends today is Agentic AI, which leverages LLMs to operate multiple tools and provide automatic configuration. In the domain of cloud computing, Agentic AI might be used, for example, with the generation of Kubernetes manifests -- structured configuration files that define containerized environments. However, effectively applying LLMs to domain-specific tasks often reveals knowledge gaps that impact the accuracy and reliability of the generated output. To address these challenges, we propose KGen, a pipeline for generating K8s manifests directly from user-described intents expressed in natural language using LLMs. Our approach leverages an extensive n-shot learning analysis to choose the appropriate number of examples that can better guide the adopted models in generating the manuscripts while also looking at the computational cost. Our results validate the use of LLM in this task and show that (as expected) increasing the number of n-shot examples can improve the quality of the generated configurations when adopting more specialized models, such as Mixtral-8x7B (which uses the Mixture of Experts approach) and Prometheus-8x7B-v2.0, but (surprisingly) for more general-purpose models like Llama3-8B and Llama3-70B, it can lead to smaller number of valid K8s manifests. These results underscore the complexities of adapting LLMs for domain-specific structured generation and emphasize the need for an in-depth analysis to determine the most effective setup, also suggesting that smaller models sometimes outperform their larger counterparts for each domain-specific task.

LogicQA: Logical Anomaly Detection with Vision Language Model Generated Questions

Yejin Kwon, Daeun Moon, Youngje Oh, Hyunsoo Yoon

Anomaly Detection (AD) focuses on detecting samples that differ from the standard pattern, making it a vital tool in process control. Logical anomalies may appear visually normal yet violate predefined constraints on object presence, arrangement, or quantity, depending on reasoning and explainability. We introduce LogicQA, a framework that enhances AD by providing industrial operators with explanations for logical anomalies. LogicQA compiles automatically generated questions into a checklist and collects responses to identify violations of logical constraints. LogicQA is training-free, annotation-free, and operates in a few-shot setting. We achieve state-of-the-art (SOTA) Logical AD performance on public benchmarks, MVTec LOCO AD, with an AUROC of 87.6% and an F1-max of 87.0% along with the explanations of anomalies. Also, our approach has shown outstanding performance on semiconductor SEM corporate data, further validating its effectiveness in industrial applications.

Overlapping Context with Variable-Length Stride Increases Diversity when Training Large Language Model for Code

Geonmo Gu, Jaeho Kwak, Haksoo Moon, Hyun Seung Shim, Yu Jin Kim, Byoungjip Kim, Moontae Lee, Hyejeong Jeon

The pretraining of code LLMs typically begins with general data and progresses to domain specific data through sequential stages. In thelatter stages, a challenging issue is that the data of a target domain can be limited in size, and conventional approach of increasing the number of epochs does not lead to a performance gain. In this paper, we propose a novel packing method, which is extracting overlapping contexts from the training data using variable length stride. Our method can mitigate the data scarcity issue by providing more diverse and abundant examples of next token prediction than non-overlapping contexts. While the training time of our approach is increased proportionally to the amount of augmented examples, we present space-efficient implementations to store overlapping contexts. Extensive experiments with real datasets show that our approach outperforms the conventional approach of controlling the number of epochs in terms of the pass@k rate.

Grammar-Constrained Decoding Makes Large Language Models Better Logical Parsers

Federico Raspanti, Tanir Özçelebi, Mike Holenderski

Large Language Models (LLMs) have shown capabilities in various natural language processing tasks, yet they often struggle with logical reasoning, particularly when dealing with complex natural language statements. To address this challenge, approaches that combine LLMs with symbolic reasoners have been proposed, where the LLM translates the natural language statements into symbolic representations, which are then verified by an external symbolic solver. However, ensuring syntactic correctness in these translations remains a significant challenge. To address this, we propose to constrain the outputs of the LLMs using Grammar-Constrained Decoding, showing that it consistently improves both syntactic correctness and semantic accuracy in logical parsing tasks. Our findings suggest that grammar constraints can serve as an effective substitute for in-context examples, especially beneficial for resource-constrained applications using smaller models.

AUTOSUMM: A Comprehensive Framework for LLM-Based Conversation Summarization

Abhinav Gupta, Devendra Singh, Greig Cowan, N Kadhiresan, Siddharth Srivastava, Sriraja Yagneswaran, Yoages Kumar Mantri

We present AUTOSUMM, a large language model (LLM)-based summarization system deployed in a regulated banking environment to generate accurate, privacy-compliant summaries of customer-advisor conversations. The system addresses challenges unique to this domain, including speaker attribution errors, hallucination risks, and short or low-information transcripts. Our architecture integrates dynamic transcript segmentation, thematic coverage tracking, and a domain specific multi-layered hallucination detection module that combines syntactic, semantic, and entailment-based checks. Human-in-the-loop feedback from over 300 advisors supports continuous refinement and auditability. Empirically, AUTOSUMM achieves a 94% factual consistency rate and a significant reduction in hallucination rate. In production, 89% of summaries required no edits, and only 1% required major corrections. A structured model version management pipeline ensures stable upgrades with minimal disruption. We detail our deployment methodology, monitoring strategy, and ethical safeguards, showing how LLMs can be reliably integrated into high-stakes, regulated workflows.

RedactOR: An LLM-Powered Framework for Automatic Clinical Data De-Identification

Praphul Singh, Charlotte Dzialo, Jangwon Kim, Sumana Srivatsa, Irfan Bulu, Sri Gadde, Krishnaram Kenthapadi

Ensuring clinical data privacy while preserving utility is critical for AI-driven healthcare and data analytics. Existing de-identification (De-ID) methods, including rule-based techniques, deep learning models, and large language models (LLMs), often suffer from recall errors, limited generalization, and inefficiencies, limiting their real-world applicability. We propose a fully automated, multi-modal framework, RedactOR for de-identifying structured and unstructured electronic health records, including clinical audio records. Our framework employs cost-efficient De-ID strategies, including intelligent routing, hybrid rule and LLM based approaches, and a two-step audio redaction approach. We present a retrieval-based entity relexicalization approach to ensure consistent substitutions of protected entities, thereby enhancing data coherence for downstream applications. We discuss key design desiderata, de-identification and relexicalization methodology, and modular architecture of RedactOR and its integration with Oracle Health Clinical AI system. Evaluated on the i2b2 2014 De-ID dataset using standard metrics with strict recall, our approach achieves competitive performance while optimizing token usage to reduce LLM costs. Finally, we discuss key lessons and insights from deployment in real-world AI-driven healthcare data pipelines.

QUPID: Quantified Understanding for Enhanced Performance, Insights, and Decisions in Korean Search Engines

Ohjoon kwon, Changsu Lee, Jihye Back, Lim Sun Suk, Inho Kang, Donghyeon Jeon

Large language models (LLMs) have been widely used for relevance assessment in information retrieval. However, our study demonstrates that combining two distinct small language models (SLMs) with different architectures can outperform LLMs in this task. Our approach—QUPID—integrates a generative SLM with an embedding-based SLM, achieving higher relevance judgment accuracy while reducing computational costs compared to state-of-the-art LLM solutions. This computational efficiency makes QUPID highly scalable for real-world search systems processing millions of queries daily. In experiments across diverse document types, our method demonstrated consistent performance improvements (Cohen's Kappa of 0.646 versus 0.387 for leading LLMs) while offering 60x faster inference times. Furthermore, when integrated into production search pipelines, QUPID improved nDCG@5 scores by 1.9\%. These findings underscore how architectural diversity in model combinations can significantly enhance both search relevance and operational efficiency in information retrieval systems.

Rethinking the Roles of Large Language Models in Chinese Grammatical Error Correction

Yinghui Li, Shang Qin, Jingheng Ye, Haojing Huang, Yangning Li, Shu-Yu Guo, Libo Qin, Xuming Hu, Wenhao Jiang, Hai-Tao Zheng, Philip S. Yu

Recently, Large Language Models (LLMs) have been widely studied by researchers for their roles in various downstream NLP tasks. As a fundamental task in the NLP field, Chinese Grammatical Error Correction (CGEC) aims to correct all potential grammatical errors in the input sentences. Previous studies have shown that LLMs' performance as correctors on CGEC remains unsatisfactory due to the challenging nature of the task. To promote the CGEC field to better adapt to the era of LLMs, we rethink the roles of LLMs in the CGEC task so that they can be better utilized and explored in CGEC. Considering the rich grammatical knowledge stored in LLMs and their powerful semantic understanding capabilities, we utilize LLMs as explainers to provide explanation information to the CGEC small models during error correction, aiming to enhance performance. We also use LLMs as evaluators to bring more reasonable CGEC evaluations, thus alleviating the troubles caused by the subjectivity of the CGEC task. In particular, our work is also an active exploration of how LLMs and small models better collaborate in downstream tasks. Extensive experiment and detailed analyses on widely used datasets verify the effectiveness of our intuition and the proposed methods.

Accelerating Antibiotic Discovery with Large Language Models and Knowledge Graphs

Maxime Delmas, Magdalena Wysocka, Danilo Gusicuma, Andre Freitas

The discovery of novel antibiotics is critical to address the growing antimicrobial resistance (AMR). However, pharmaceutical industries face high costs (over 1 billion), long timelines, and a high failure rate, worsened by the rediscovery of known compounds. We propose an LLM-based pipeline that acts as an alert system, detecting prior evidence of antibiotic activity to prevent costly rediscoveries. The system integrates literature on organisms and chemicals into a Knowledge Graph (KG), ensuring taxonomic resolution, synonym handling, and multi-level evidence classification. We tested the pipeline on a private list of 73 potential antibiotic-producing organisms, disclosing 12 negative hits for evaluation. The results highlight the effectiveness of the pipeline for evidence reviewing, reducing false negatives, and accelerating decision-making. The KG for negative hits as well as the user interface for interactive exploration are available at https://github.com/idiap/abroad-kg-store and https://github.com/idiap/abroad-demo-webapp.

FoodTaxo: Generating Food Taxonomies with Large Language Models

Pascal Wullschleger, Majid Zarharan, Donnacha Daly, Marc Pouly, Jennifer Foster

We investigate the utility of Large Language Models for automated taxonomy generation and completion specifically applied to taxonomies from the food technology industry. We explore the extent to which taxonomies can be completed from a seed taxonomy or generated without a seed from a set of known concepts, in an iterative fashion using recent prompting techniques. Experiments on five taxonomies using an open-source LLM (Llama-3), while promising, point to the difficulty of correctly placing inner nodes.

Enriching children's stories with LLMs: Delivering multilingual data enrichment for children's books at scale and across markets

Zarah Weiss, Christof Meyer, Mikael Andersson

This paper presents a user-centered, empirically guided approach to multilingual metadata enrichment for children's books. We combine LLMs with human-in-the-loop quality control in a scalable CI/CD pipeline to curate brand collections that enhance book discovery and engagement for young readers across multiple European markets. Our results demonstrate that this hybrid approach delivers high-quality, child-appropriate labels, improves user experience, and accelerates deployment in real-world production environments. This work offers practical insights for applying generative NLP in the media and publishing industry.

Advanced Messaging Platform (AMP): Pipeline for Automated Enterprise Email Processing

Simerjot Kaur, Charese Smiley, Keshav Ramani, Elena Kochkina, Mathieu Sibue, Samuel Mensah, Pietro Totis, Cecilia Tilli, Toyin Aguda, Daniel Borrajo, Manuela Veloso

Understanding and effectively responding to email communication remains a critical yet complex challenge for current AI techniques, especially in corporate environments. These tasks are further complicated by the need for domain-specific knowledge, accurate entity recognition, and high precision to prevent costly errors. While recent advances in AI, specifically Large Language Models (LLMs), have made strides in natural language understanding, they often lack business-specific expertise required in such settings. In this work, we present Advanced Messaging Platform (AMP), a production-grade AI pipeline that automates email response generation at scale in real-world enterprise settings. AMP has been in production for more than a year, processing thousands of emails daily while maintaining high accuracy and adaptability to evolving business needs.

Semantic Outlier Removal with Embedding Models and LLMs

M. Eren Akbiyik, João Almeida, Rik Melis, Ritu Sriram, Viviana Petrescu, Vilhjálmur Vilhjálmsson

Modern text processing pipelines demand robust methods to remove extraneous content while preserving a document's core message. Traditional approaches—such as HTML boilerplate extraction or keyword filters—often fail in multilingual settings and struggle with context-sensitive nuances, whereas Large Language Models (LLMs) offer improved quality at high computational cost. We introduce SORE (Semantic Outlier Removal), a cost-effective, transparent method that leverages multilingual sentence embeddings and approximate nearest-neighbor search to identify and excise unwanted text segments. By first identifying core content via metadata embedding and then flagging segments that either closely match predefined outlier groups or deviate significantly from the core, SORE achieves near-LLM extraction precision at a fraction of the cost. Experiments on HTML datasets demonstrate that SORE outperforms structural methods and yield high precision in diverse scenarios. Our system is currently deployed in production, processing millions of documents daily across multiple languages while maintaining both efficiency and accuracy. To facilitate reproducibility and further research, we will publicly release our implementation and evaluation datasets.

SLENDER: Structured Outputs for SLM-based NER in Low-Resource Englishes

Nicole Ren, James Teo

Named Entity Recognition (NER) for low-resource variants of English remains challenging, as most NER models are trained on datasets predominantly focused on American or British English. While recent work has shown that proprietary Large Language Models (LLMs) can perform NER effectively in low-resource settings through in-context learning, practical deployment is limited by their high computational costs and privacy concerns. Open-source Small Language Models (SLMs) offer promising alternatives, but the tendency of these Language Models (LM) to hallucinate poses challenges for production use. To address this, we introduce SLENDER, a novel output format for LM-based NER that achieves a three-fold reduction in inference time on average compared to JSON format, which is widely used for structured outputs. Our approach using Gemma-2-9B-it with the SLENDER output format and constrained decoding in zero-shot settings outperforms the en_core_web_trf model from SpaCy, an industry-standard NER tool, in all five regions of the Worldwide test set.

A Large-Scale Real-World Evaluation of an LLM-Based Virtual Teaching Assistant

Sunjun Kweon, Sooyohn Nam, Hyunseung Lim, Hwajung Hong, Edward Choi

Virtual Teaching Assistants (VTAs) powered by Large Language Models (LLMs) have the potential to enhance student learning by providing instant feedback and facilitating multi-turn interactions. However, empirical studies on their effectiveness and acceptance in real-world classrooms are limited, leaving their practical impact uncertain. In this study, we develop an LLM-based VTA and deploy it in an introductory AI programming course with 477 graduate students. To assess how student perceptions of the VTA's performance evolve over time, we conduct three rounds of comprehensive surveys at different stages of the course. Additionally, we analyze 3,869 student–VTA interaction pairs to identify common question types and engagement patterns. We then compare these interactions with traditional student-human instructor interactions to evaluate the VTA's role in the learning process. Through a large-scale empirical study and interaction analysis, we assess the feasibility of deploying VTAs in real-world classrooms and identify key challenges for broader adoption. Finally, we release the source code of our VTA system, fostering future advancements in AI-driven education.

Filter-And-Refine: A MLLM Based Cascade System for Industrial-Scale Video Content Moderation

Zixuan Wang, jinghao shi, Hanzhong Liang, Xiang Shen, Vera Wen, Zhiqian Chen, Yifan Wu, Zhixin Zhang, Hongyu Xiong

Effective content moderation is essential for video platforms to safeguard user experience and uphold community standards. While traditional video classification models effectively handle well-defined moderation tasks, they struggle with complicated scenarios such as implicit harmful content and contextual ambiguity. Multimodal large language models (MLLMs) offer a promising solution to these limitations with their superior cross-modal reasoning and contextual understanding. However, two key challenges hinder their industrial adoption. First, the high computational cost of MLLMs makes full-scale deployment impractical. Second, adapting generative models for discriminative classification remains an open research problem. In this paper, we first introduce an efficient method to transform a generative MLLM into a multimodal classifier using minimal discriminative training data. To enable industry-scale deployment, we then propose a router-ranking cascade system that integrates MLLMs with a lightweight router model. Offline experiments demonstrate that our MLLM-based approach improves F1 score by 66.50% over traditional classifiers while requiring only 2% of the fine-tuning data. Online evaluations show that our system increases automatic content moderation volume by 41%, while the cascading deployment reduces computational cost to only 1.5% of direct full-scale deployment.

ASK: Aspects and Retrieval based Hybrid Clarification in Task Oriented Dialogue Systems

Rishav Sahay, Lavanya Sita Tekumalla, Purav Aggarwal, Arihant Jain, Anoop Saladi

Ambiguous user queries pose a significant challenge in task-oriented dialogue systems relying on information retrieval. While Large Language Models (LLMs) have shown promise in generating clarification questions to tackle query ambiguity, they rely solely on the top-k retrieved documents for clarification which fails when ambiguity is too high to retrieve relevant documents in the first place. Traditional approaches lack principled mechanisms to determine when to use broad domain knowledge vs specific retrieved document context for clarification. We propose AsK, a novel hybrid approach that dynamically chooses between document-based or aspect-based clarification based on query ambiguity. Our approach requires no labeled clarification data and introduces: (1) Weakly-supervised Longformer-based ambiguity analysis, (2) Automated domain-specific aspect generation using clustering and LLMs and (3) LLM-powered clarification generation. AsK demonstrates significant improvements over baselines in both single-turn and multi-turn settings (recall@5 gain of 20%) when evaluated on product troubleshooting and product search datasets.

A Framework for Flexible Extraction of Clinical Event Contextual Properties from Electronic Health Records

Shubham Agarwal, Thomas Searle, Mart Ratas, Anthony Shek, James Teo, Richard Dobson

Electronic Health Records contain vast amounts of valuable clinical data, much of which is stored as unstructured text. Extracting meaningful clinical events (e.g., disorders, symptoms, findings, medications, and procedures etc.) in context within real-world healthcare settings is crucial for enabling downstream applications such as disease prediction, clinical coding for billing and decision support. After Named Entity Recognition and Linking (NER+L) methodology, the identified concepts need to be further classified (i.e. contextualized) for distinct properties such as their relevance to the patient, their temporal and negated status for meaningful clinical use. We present a solution that, using an existing NER+L approach - MedCAT, classifies and contextualizes medical entities at scale. We evaluate the NLP approaches through 14 distinct real-world clinical text classification projects, testing our suite of models tailored to different clinical NLP needs. For tasks requiring high minority class recall, BERT proves the most effective when coupled with class imbalance mitigation techniques, outperforming Bi-LSTM with up to 28%. For majority class focused tasks, Bi-LSTM offers a lightweight alternative with, on average, 32% faster training time and lower computational cost. Importantly, these tools are integrated into an openly available library, enabling users to select the best model for their specific downstream applications.

Small Language Models in the Real World: Insights from Industrial Text Classification

Lujun LI, Lama sleem, Niccolo' GENTILE, Geoffrey NICHIL, Radu State

With the emergence of ChatGPT, Transformer models have significantly advanced text classification and related tasks. Decoder-only models such as Llama exhibit strong performance and flexibility, yet they suffer from inefficiency on inference due to token-by-token generation, and their effectiveness in text classification tasks heavily depends on prompt quality. Moreover, their substantial GPU resource requirements often limit widespread adoption. Thus, the question of whether smaller language models are capable of effectively handling text classification tasks emerges as a topic of significant interest. However, the selection of appropriate models and methodologies remains largely underexplored. In this paper, we conduct a comprehensive evaluation of prompt engineering and supervised fine-tuning methods for transformer-based text classification. Specifically, we focus on practical industrial scenarios, including email classification, legal document categorization, and the classification of extremely long academic texts. We examine the strengths and limitations of smaller models, with particular attention to both their performance and their efficiency in Video Random-Access Memory (VRAM) utilization, thereby providing valuable insights for the local deployment and application of compact models in industrial settings.

AutoChunker: Structured Text Chunking and its Evaluation

Arihant Jain, Purav Aggarwal, Anoop Saladi

Text chunking is fundamental to modern retrieval-augmented systems, yet existing methods often struggle with maintaining semantic coherence, both within and across chunks, while dealing with document structure and noise. We present AutoChunker, a bottom-up approach for text chunking that combines document structure awareness with noise elimination. AutoChunker leverages language models to identify and segregate logical units of information (a chunk) while preserving document hierarchy through a tree-based representation. To evaluate the chunking operator, we introduce a comprehensive evaluation framework based on five core tenets: noise reduction, completeness, context coherence, task relevance, and retrieval performance. Experimental results on Support and Wikipedia articles demonstrate that AutoChunker significantly outperforms existing methods, reducing noise while improving chunk completeness compared to state-of-the-art baselines. When integrated with an online product support system, our approach led to improvements in retrieval performance and customer return rates. Our work not only advances the state of text chunking but also provides a standardized framework for evaluating chunking strategies, addressing a critical gap in the field.

SQLGenie: A Practical LLM based System for Reliable and Efficient SQL Generation

Pushpendu Ghosh, Aryan Jain, Promod Yenigalla

Large Language Models (LLMs) enable natural language to SQL conversion, allowing users to query databases without SQL expertise. However, generating accurate, efficient queries is challenging due to ambiguous intent, domain knowledge requirements, and database constraints. Extensive reasoning improves SQL quality but increases computational costs and latency. We propose SQLGenie, a practical system for reliable SQL generation. It consists of three components: (1) \textbf{Table Onboarder}, which analyzes new tables, optimizes indexing, partitions data, identifies foreign key relationships, and stores schema details for SQL generation; (2) \textbf{SQL Generator}, an LLM-based system producing accurate SQL; and (3) \textbf{Feedback Augmentation}, which filters correct query-SQL pairs, leverages multiple LLM agents for complex SQL, and stores verified examples. SQLGenie achieves state-of-the-art performance on public benchmarks (92.8\% execution accuracy on WikiSQL, 82.1\% of Spider, 73.8\% on BIRD) and internal datasets, surpassing the best single-LLM baseline by 21.5\% and the strongest pipeline competitor by 5.3\%. Its hybrid variant optimally balances accuracy and efficiency, reducing generation time by 64\% compared to traditional multi-LLM approaches while maintaining competitive accuracy.

Hard Negative Mining for Domain-Specific Retrieval in Enterprise Systems

Hansa Meghwani, Amit Agarwal, Priyaranjan Pattnayak, Hitesh Laxmichand Patel, Srikant Panda

Enterprise search systems often struggle to retrieve accurate, domain-specific information due to semantic mismatches and overlapping terminologies. These issues can degrade the performance of downstream applications such as knowledge management, customer support, and retrieval-augmented generation agents. To address this challenge, we propose a scalable hard-negative mining framework tailored specifically for domain-specific enterprise data. Our approach dynamically selects semantically challenging but contextually irrelevant documents to enhance deployed re-ranking models. Our method integrates diverse embedding models, performs dimensionality reduction, and uniquely selects hard negatives, ensuring computational efficiency and semantic precision. Evaluation on our proprietary enterprise corpus (cloud services domain) demonstrates substantial improvements of 15\% in MRR@3 and 19\% in MRR@10 compared to state-of-the-art baselines and other negative sampling techniques. Further validation on public domain-specific datasets (FiQA, Climate Fever, TechQA) confirms our method's generalizability and readiness for real-world applications.

Interpretable Company Similarity with Sparse Autoencoders

Marco Molinari, Victor Shao, Luca Imeneo, Mateusz Mikolajczak, Abhimanyu Pandey, Vladimir Tregubiak, Sebastião Kuznetsov Ryder Torres Pereira

Determining company similarity is a vital task in finance, underpinning risk management, hedging, and portfolio diversification. Practitioners often rely on sector and industry classifications such as SIC and GICS codes to gauge similarity, the former being used by the U.S. Securities and Exchange Commission (SEC), and the latter widely used by the investment community. Since these classifications lack granularity and need regular updating, using clusters of embeddings of company descriptions has been proposed as a potential alternative, but the lack of interpretability in token embeddings poses a significant barrier to adoption in high-stakes contexts. Sparse Autoencoders (SAEs) have shown promise in enhancing the interpretability of Large Language Models (LLMs) by decomposing Large Language Model (LLM) activations into interpretable features. Moreover, SAEs capture an LLM's internal representation of a company description, as opposed to semantic similarity alone, as is the case with embeddings. We apply SAEs to company descriptions, and obtain meaningful clusters of equities. We benchmark SAE features against SIC-codes, Industry codes, and Embeddings. Our results demonstrate that SAE features surpass sector classifications and embeddings in capturing fundamental company characteristics. This is evidenced by their superior performance in correlating logged monthly returns – a proxy for similarity – and generating higher Sharpe ratios in co-integration trading strategies, which underscores deeper fundamental similarities among companies. Finally, we verify the interpretability of our clusters, and demonstrate that sparse features form simple and interpretable explanations for our clusters.

Domain Adaptation of Foundation LLMs for e-Commerce

Christian Herold, Michael Kozielski, Tala Bazazo, Pavel Petrushkov, Yannick Versley, Seyyed Hadi Hashemi, Patrycja Cieplicka, Dominika Basaj, Shahram Khadivi

We present the e-Llama models: 8 billion and 70 billion parameter large language models that are adapted towards the e-commerce domain. These models are meant as foundation models with deep knowledge about e-commerce, that form a base for instruction- and fine-tuning. The e-Llama models are obtained by continuously pretraining the Llama 3.1 base models on 1 trillion tokens of domain-specific data. We discuss our approach and motivate our choice of hyperparameters with a series of ablation studies. To quantify how well the models have been adapted to the e-commerce domain, we define and implement a set of multilingual, e-commerce specific evaluation tasks. We show that, when carefully choosing the training setup, the Llama 3.1 models can be adapted towards the new domain without sacrificing significant performance on general domain tasks. We also explore the possibility of merging the adapted model and the base model for a better control of the performance trade-off between domains.

sudo rm -rf agentic_security

Sejin Lee, Jian Kim, Haon Park, Ashkan Yousefpour, Sangyoon Yu, Min Song

Large Language Models (LLMs) are increasingly deployed as computer-use agents, autonomously performing tasks within real desktop or web environments. While this evolution greatly expands practical use cases for humans, it also creates serious security exposures. We present SUDO (Screen-based Universal Detox2tox Offense), a novel attack framework that systematically bypasses refusal-trained safeguards in commercial computer-use agents, such as Claude for Computer Use. The core mechanism, Detox2tox, transforms harmful requests (that agents initially reject) into seemingly benign requests via detoxification, secures detailed instructions from advanced vision language models (VLMs), and then reintroduces malicious content via toxification just before execution. Unlike conventional jailbreaks, SUDO iteratively refines its attacks based on a built-in refusal feedback, making it increasingly effective against robust policy filters. In extensive tests spanning 50 real-world tasks and multiple state-of-the-art VLMs, SUDO achieves a stark attack success rate of 24.41\% (with no refinement), and up to 41.33\% (by its iterative refinement) in Claude for Computer Use. By revealing these vulnerabilities and demonstrating the ease with which they can be exploited in real-world computing environments, this paper highlights an immediate need for robust, context-aware safeguards. WARNING: This paper includes harmful or offensive model outputs.

MultiMed: Multilingual Medical Speech Recognition via Attention Encoder Decoder

Khai Le-Duc, Phuc Phan, Tan-Hanh Pham, Bach Phan Tat, Minh-Huong Ngo, Thanh Nguyen-Tang, Truong-Son Hy

Multilingual automatic speech recognition (ASR) in the medical domain serves as a foundational task for various downstream applications such as speech translation, spoken language understanding, and voice-activated assistants. This technology improves patient care by enabling efficient communication across language barriers, alleviating specialized workforce shortages, and facilitating improved diagnosis and treatment, particularly during pandemics. In this work, we introduce MultiMed, the first multilingual medical ASR dataset, along with the first collection of small-to-large end-to-end medical ASR models, spanning five languages: Vietnamese, English, German, French, and Mandarin Chinese. To our best knowledge, MultiMed stands as the world's largest medical ASR dataset across all major benchmarks: total duration, number of recording conditions, number of accents, and number of speaking roles. Furthermore, we present the first multilinguality study for medical ASR, which includes reproducible empirical baselines, a monolinguality-multilinguality analysis, Attention Encoder Decoder (AED) vs Hybrid comparative study and a linguistic analysis. We present practical ASR end-to-end training schemes optimized for a fixed number of trainable parameters that are common in industry settings. All code, data, and models are available online.

MICE: Mixture of Image Captioning Experts Augmented e-Commerce Product Attribute Value Extraction

Jiaying Gong, Hongda Shen, Janet Jenq

Attribute value extraction plays a crucial role in enhancing e-commerce search, filtering, and recommendation systems. However, prior visual attribute value extraction methods typically rely on both product images and textual information such as product descriptions and titles. In practice, text can be ambiguous, inaccurate, or unavailable, which can degrade model performance. We propose Mixture of Image Captioning Experts (MICE), a novel augmentation framework for product attribute value extraction. MICE leverages a curated pool of image captioning models to generate accurate captions from product images, resulting in robust attribute extraction solely from an image. Extensive experiments on the public ImplicitAVE dataset and a proprietary women's tops dataset demonstrate that MICE significantly improves the performance of state-of-the-art large multimodal models (LMMs) in both zero-shot and fine-tuning settings. An ablation study validates the contribution of each component in the framework. MICE's modular design offers scalability and adaptability, making it well-suited for diverse industrial applications with varying computational and latency requirements.

FINKRX: Establishing Best Practices for Korean Financial NLP

Guijin Son, Hyunwoo Ko, Hanearl Jung, Chami Hwang

In this work, we present the first open leaderboard for evaluating Korean large language models focused on finance. Operated for about eight weeks, the leaderboard evaluated 1,119 submissions on a closed benchmark covering five MCQA categories: finance and accounting, stock price prediction, domestic company analysis, financial markets, and financial agent tasks and one open-ended qa task. Building on insights from these evaluations, we release an open instruction dataset of 80k instances and summarize widely used training strategies observed among top-performing models. Finally, we introduce FINKRX, a fully open and transparent LLM built using these best practices. We hope our contributions help advance the development of better and safer financial LLMs for Korean and other languages.

Judging the Judges: Can Large Vision-Language Models Fairly Evaluate Chart Comprehension and Reasoning?

Md Tahmid Rahman Laskar, Mohammed Saidul Islam, Ridwan Mahbub, Ahmed Masry, Mizanur Rahman, Amran Bhuiyan, Mir Tafseer Nayeem, Shafiq Joty, Enamul Hoque, Jimmy Huang

Charts are ubiquitous as they help people understand and reason with data. Recently, various downstream tasks, such as chart question answering, chart2text, and fact-checking, have emerged. Large Vision-Language Models (LVLMs) show promise in tackling these tasks, but their evaluation is costly and time-consuming, limiting real-world deployment. While using LVLMs as judges to assess chart comprehension capabilities of other LVLMs could streamline evaluation processes, challenges like proprietary datasets, restricted access to powerful models, and evaluation costs hinder their adoption in industrial settings. To this end, we present a comprehensive evaluation of 13 open-source LVLMs as judges for diverse chart comprehension and reasoning tasks. We design both pairwise and pointwise evaluation tasks covering criteria like factual correctness, informativeness, and relevancy. Additionally, we analyze LVLM judges based on format adherence, positional consistency, length bias, and instruction-following. We focus on cost-effective LVLMs (<10B parameters) suitable for both research and commercial use, following a standardized evaluation protocol and rubric to measure the LVLM judge accuracy. Experimental results reveal notable variability: while some open LVLM judges achieve GPT-4-level evaluation performance (about 80\% agreement with GPT-4 judgments), others struggle (below 10\% agreement). Our findings highlight that state-of-the-art open-source LVLMs can serve as cost-effective automatic evaluators for chart-related tasks, though biases such as positional preference and length bias persist.

One Missing Piece for Open-Source Reasoning Models: A Dataset to Mitigate Cold-Starting Short CoT LLMs in RL

Hyungjoo Chae, Dongjin Kang, Jihyuk Kim, Beong-woo Kwak, Sunghyun Park, Haeju Park, Jinyoung Yeo, Moontae Lee, Kyungjae Lee

With the release of R1, a publicly available large reasoning model (LRM), researchers commonly train new LRMs by training language models on R1's long chain-of-thought (CoT) inferences. While prior works show that LRMs' capabilities can be reproduced through direct distillation, the continued reliance on the existing models (e.g., R1) remains a critical limitation in advancing the field. As a first step toward independent LRM development, this paper explores the possibility of constructing a long CoT dataset with LLMs that are not trained for inference-time scaling. To this end, we present the Long CoT Collection, a dataset of 100K CoT rationales annotated using existing short CoT LLMs. We develop a pipeline that induces o1's novel reasoning strategies into short CoT LLMs, enabling them to think longer and introducing controllability over the thought budget to better manage the overthinking problem. Our extensive analyses validate that our dataset achieves quality comparable to—or slightly below—R1. Furthermore, our experiments demonstrate that training on our dataset not only strengthens general reasoning skills, but also provides a strong foundation for reinforcement learning—models initialized on our data achieve 2-3x larger gains with RLVR. We make the codes, datasets, and models publicly available at LINK.

Unifying Streaming and Non-streaming Zipformer-based ASR

Bidisha Sharma, Karthik Pandia Durai, Shankar Venkatesan, Jeena J Prakash, Shashi Kumar, Malolan Chetlur, Andreas Stolcke

There has been increasing interest in unifying streaming and non-streaming automatic speech recognition (ASR) models to reduce development, training, and deployment costs. We present a unified framework that trains a single end-to-end ASR model for both streaming and non-streaming applications, leveraging future context information. We propose to use dynamic right-context through the chunked attention masking in the training of zipformer-based ASR models. We demonstrate that using right-context is more effective in zipformer models compared to other conformer models due to its multi-scale nature. We analyze the effect of varying the number of right-context frames on accuracy and latency of the streaming ASR models. We use Librispeech and large in-house conversational datasets to train different versions of streaming and non-streaming models and evaluate them in a production grade server-client setup across diverse testsets of different domains. The proposed strategy reduces word error by relative 7.9% with a small degradation in user-perceived latency. By adding more right-context frames, we are able to achieve streaming performance close to that of non-streaming models. Our approach also allows flexible control of the latency-accuracy tradeoff according to customers requirements.

A Semi-supervised Scalable Unified Framework for E-commerce Query Classification

Chunyuan Yuan, Chong Zhang, Zheng Fang, Ming Pang, Xue Jiang, Changping Peng, Zhangang Lin, Ching Law

Query classification, including multiple subtasks such as intent and category prediction, is a vital part of e-commerce applications. E-commerce queries are usually short and lack context, and the information between labels cannot be used, resulting in insufficient prior information for modeling. Most existing industrial query classification methods rely on users' posterior click behavior to construct training samples, resulting in a Matthew vicious cycle. Furthermore, the subtasks of query classification lack a unified framework, leading to low efficiency for algorithm improvement. In this paper, we propose a novel Semi-supervised Scalable Unified Framework (SSUF), containing multiple enhanced modules to unify the query classification tasks. The knowledge-enhanced module uses world knowledge to enhance query representations and solve the problem of insufficient query information. The label-enhanced module uses label semantics and semi-supervised signals to reduce the dependence on posterior labels. The structure-enhanced module enhances the label representation based on the complex label relations. Each module is highly pluggable, and input features can be added or removed as needed according to each subtask. We conduct extensive offline and online A/B experiments, and the results show that SSUF significantly outperforms the state-of-the-art models.

BI-Bench : A Comprehensive Benchmark Dataset and Unsupervised Evaluation for BI Systems

Ankush Gupta, Aniya Aggarwal, Shivangi Bithel, Arvind Agarwal

A comprehensive benchmark is crucial for evaluating automated Business Intelligence (BI) systems and their real-world effectiveness. We propose BI-Bench, a holistic, end-to-end benchmarking framework that assesses BI systems based on the quality, relevance, and depth of insights. It categorizes queries into descriptive, diagnostic, predictive, and prescriptive types, aligning with practical BI needs. Our fully automated approach enables custom benchmark generation tailored to specific datasets. Additionally, we introduce an automated evaluation mechanism within BI-Bench that removes reliance on strict ground truth, ensuring scalable and adaptable assessments. By addressing key limitations, it offers a flexible and robust, user-centered methodology for advancing next-generation BI systems.

Consistency-Aware Online Multi-Objective Alignment for Related Search Query Generation

Shuxian Bi, Chongming Gao, Wenjie Wang, Yueqi Mou, Chenxu Wang, Tang Biao, Peng Yan, Fuli Feng

Modern digital platforms rely on related search query recommendations to enhance engagement, yet existing methods fail to reconcile click-through rate (CTR) optimization with topic expansion. We propose CMAQ, a Consistent Multi-Objective Aligned Query generation framework that harmonizes these goals through three components: (1) reward modeling to quantify objectives, (2) style alignment for format compliance, and (3) consistency-aware optimization to coordinate joint improvements. CMAQ employs adaptive-beta scaled DPO with geometric mean rewards, balancing CTR and expansion while mitigating objective conflicts. Extensive offline and online evaluations in a large-scale industrial setting demonstrate CMAQ's superiority, achieving significant CTR gains (+2.3%) and higher human-rated query quality compared to state-of-the-art methods. Our approach enables high-quality query generation while sustaining user engagement and platform ecosystem health.

Towards Generating Controllable and Solvable Geometry Problem by Leveraging Symbolic Deduction Engine

Zhuoxuan Jiang, Tianyang Zhang, Peiyan Peng, JING CHEN, Yinong Xun, Haotian Zhang, Lichi Li, Yong Li, Shaohua Zhang

Generating high-quality geometry problems is both an important and challenging task in education. Compared to math word problems, geometry problems further emphasize multi-modal formats and the translation between informal and formal languages. In this paper, we introduce a novel task for geometry problem generation and propose a new pipeline method: the Symbolic Deduction Engine-based Geometry Problem Generation framework (SDE-GPG). The framework leverages a symbolic deduction engine and contains four main steps: (1) searching a predefined mapping table from knowledge points to extended definitions, (2) sampling extended definitions and performing symbolic deduction, (3) filtering out unqualified problems, and (4) generating textual problems and diagrams. Specifically, our method supports to avoid inherent biases in translating natural language into formal language by designing the mapping table, and guarantees to control the generated problems in terms of knowledge points and difficulties by an elaborate checking function. With obtained formal problems, they are translated to natural language and the accompanying diagrams are automatically drew by rule-based methods. We conduct experiments using real-world combinations of knowledge points from two public datasets. The results demonstrate that the SDE-GPG can effectively generate readable, solvable and controllable geometry problems.

TableCoder: Table Extraction from Text via Reliable Code Generation

Haoyu Dong, Yue Hu, Huailiang Peng, Yanan Cao

This paper introduces a task aimed at extracting structured tables from text using natural language (NL) instructions. We present TableCoder, an approach that leverages the symbolic nature of code to enhance the robustness of table structure construction and content extraction. TableCoder first generates Python classes or SQL statements to explicitly construct table structures, capturing semantic ontology, computational dependencies, numerical properties, and format strings. This approach reliably mitigates issues such as structural errors, erroneous computations, and mismatched value types. Subsequently, TableCoder proposes grounded content extraction, populating table cells sequentially and maintaining the exact order in which they are mentioned in the source text. By simulating a grounded "translation" from text to code, this method reduces the likelihood of omissions and hallucinations. Experimental results demonstrate that TableCoder significantly improves F1 scores and mitigates hallucination and computational errors, crucial for high-stakes applications like government data analytics and financial compliance reporting. Moreover, the code-generation-based method naturally integrates with standard SQL databases and Python workflows, ensuring seamless deployment in existing enterprise data pipelines.

Are LLMs reliable? An exploration of the reliability of large language models in clinical note generation

Kristine Ann M. Carandang, Jasper Meynard Arana, Ethan Robert Casin, Christopher Monterola, Daniel Stanley Tan, Jesus Felix B. Valenzuela, Christian Alis

Due to the legal and ethical responsibilities of healthcare providers (HCPs) for accurate documentation and protection of patient data privacy, the natural variability in the responses of large language models (LLMs) presents challenges for incorporating clinical note generation (CNG) systems, driven by LLMs, into real-world clinical processes. The complexity is further amplified by the detailed nature of texts in CNG. To enhance the confidence of HCPs in tools powered by LLMs, this study evaluates the reliability of 12 open-weight and proprietary LLMs from Anthropic, Meta, Mistral, and OpenAI in CNG in terms of their ability to generate notes that are string equivalent (consistency rate), have the same meaning (semantic consistency) and are correct (semantic similarity), across several iterations using the same prompt. The results show that (1) LLMs from all model families are stable, such that their responses are semantically consistent despite being written in various ways, and (2) most of the LLMs generated notes close to the corresponding notes made by experts. Overall, Meta's Llama 70B was the most reliable, followed by Mistral's Small model. With these findings, we recommend the local deployment of these relatively smaller open-weight models for CNG to ensure compliance with data privacy regulations, as well as to improve the efficiency of HCPs in clinical documentation.

Convert Language Model into a Value-based Strategic Planner

Xiaoyu Wang, Yue Zhao, Qingqing Gu, Zhonglin Jiang, Yong Chen, Luo Ji

Emotional support conversation (ESC) aims to alleviate the emotional distress of individuals through effective conversations. Although large language models (LLMs) have obtained remarkable progress on ESC, most of these studies might not define the diagram from the state model perspective, therefore providing a suboptimal solution for long-term satisfaction. To address such an issue, we leverage the Q-learning on LLMs, and propose a framework called straQ*. Our framework allows a plug-and-play LLM to bootstrap the planning during ESC, determine the optimal strategy based on long-term returns, and finally guide the LLM to response. Substantial experiments on ESC datasets suggest that straQ* outperforms many baselines, including direct inference, self-refine, chain of thought, finetuning, and finite state machines.

ConCodeEval: Evaluating Large Language Models for Code Constraints in Domain-Specific Languages

Mehant Kammakomati, Sameer Pimparkhede, Srikanth G. Tamilselvam, Prince Kumar, Pushpak Bhattacharyya

System-level programming is essential for modern enterprise infrastructure, enabling the automation and management of complex systems through declarative code. Developers write this code based on schemas, which themselves are a form of code that defines constraints like data types and required fields. These schemas help ensure operational correctness and smooth integration across systems. However, as enterprise schemas become complex, manually writing code adhering to these constraints becomes challenging for developers. Large Language Models (LLMs) have demonstrated potential in code generation and natural language understanding, particularly in zero-shot and few-shot settings. However, applying LLMs to handle constraints represented in code, essential for system-level programming rather than natural language, has not been explored. Hence, we introduce ConCodeEval, a study across two key dimensions: format and constraint efficacy, with a first-of-its-kind benchmark involving two novel experiments for code constraints across five representations (JSON, YAML, XML, Python, and natural language). Our findings suggest that conscious choice of representations can lead to optimal use of LLMs in enterprise use cases involving constraints. Nonetheless, LLMs continue to struggle significantly with code constraints, motivating the need for innovation in this direction.

Unveiling Dual Quality in Product Reviews: An NLP-Based Approach

Rafał Poświata, Marcin Mirończuk, Sławomir Dadas, Małgorzata Grębowiec, Michał Wiktor Perełkiewicz

Consumers often face inconsistent product quality, particularly when identical products vary between markets, a situation known as the dual quality problem. To identify and address this issue, automated techniques are needed. This paper explores how natural language processing (NLP) can aid in detecting such discrepancies and presents the full process of developing a solution. First, we describe in detail the creation of a new Polish-language dataset with 1,957 reviews, 540 highlighting dual quality issues. We then discuss experiments with various approaches like SetFit with sentence-transformers, transformer-based encoders, and LLMs, including error analysis and robustness verification. Additionally, we evaluate multilingual transfer using a subset of opinions in English, French, and German. The paper concludes with insights on deployment and practical applications.

Enhancing Marker Scoring Accuracy through Ordinal Confidence Modelling in Educational Assessments

Abhirup Chakravarty, Mark Brenchley, Trevor Breakspear, Ian Lewin, Yan Huang

A key ethical challenge in Automated Essay Scoring (AES) is ensuring that scores are only released when they meet high reliability standards. Confidence modelling addresses this by assigning a reliability estimate measure, in the form of a confidence score, to each automated score. In this study, we frame confidence estimation as a classification task: predicting whether an AES-generated score correctly places a candidate in the appropriate CEFR level. While this is a binary decision, we leverage the inherent granularity of the scoring domain in two ways. First, we reformulate the task as an n-ary classification problem using score binning. Second, we introduce a set of novel Kernel Weighted Ordinal Categorical Cross Entropy (KWOCCE) loss functions that incorporate the ordinal structure of CEFR labels. Our best-performing model achieves an F1 score of 0.97, and enables the system to release 47 percent of scores with 100 percent CEFR agreement and 99 percent with at least 95 percent CEFR agreement – compared to 92 percent (approx.) CEFR agreement from the standalone AES model where we release all AM predicted scores.

A Practical Approach for Building Production-Grade Conversational Agents with Workflow Graphs

Chiwan Park, Wonjun Jang, Daeryong Kim, Aelim Ahn, Kichang Yang, Woosung Hwang, Jihyeon Roh, Hyerin Park, Hyosun Wang, Min Seok Kim, don.kang

The advancement of Large Language Models (LLMs) has led to significant improvements in various service domains, including search, recommendation, and chatbot applications. However, applying state-of-the-art (SOTA) research to industrial settings presents challenges, as it requires maintaining flexible conversational abilities while also strictly complying with service-specific constraints. This can be seen as two conflicting requirements due to the probabilistic nature of LLMs. In this paper, we propose our approach to addressing this challenge and detail the strategies we employed to overcome their inherent limitations in real-world applications. We conduct a practical case study of a conversational agent designed for the e-commerce domain, detailing our implementation workflow and optimizations. Our findings provide insights into bridging the gap between academic research and real-world application, introducing a framework for developing scalable, controllable, and reliable AI-driven agents.

Session 10: Machine Translation (Tuesday 29 July, 16:00-17:30, Hall 4/5)

AskQE: Question Answering as Automatic Evaluation for Machine Translation

Dayeon Ki, Kevin Duh, Marine Carpuat

How can a monolingual English speaker determine whether an automatic translation in French is good enough to be shared? Existing MT error detection and quality estimation (QE) techniques do not address this practical scenario. We introduce AskQE, a question generation and answering framework designed to detect critical MT errors and provide actionable feedback, helping users decide whether to accept or reject MT outputs even without the knowledge of the target language. We propose an optimized version of AskQE using LLaMA-3 70B given entailed facts during question generation. We evaluate our method on the ContraTICO dataset across five language pairs, and show that AskQE effectively identifies critical MT errors with high correlations with established QE metrics. We further extend our analysis on the BioMQM dataset of naturally occurring MT errors, where we show that AskQE has higher Kendall's Tau correlation and decision accuracy with human ratings compared to other QE metrics.

Going Beyond Your Expectations in Latency Metrics for Simultaneous Speech Translation

Jorge Iranzo-Sánchez, Javier Iranzo-Sánchez, Adrià Giménez, Jorge Civera

Current evaluation practices in Simultaneous Speech Translation (SimulST) systems typically involve segmenting the input audio and corresponding translations, calculating quality and latency metrics for each segment, and averaging the results. Although this approach may provide a reliable estimation of translation quality, it can lead to misleading values of latency metrics due to an inherent assumption that average latency values are good enough estimators of SimulST systems' response time. However, our detailed analysis of latency evaluations for state-of-the-art SimulST systems demonstrates that latency distributions are often skewed and subject to extreme variations. As a result, the mean in latency metrics fails to capture these anomalies, potentially masking the lack of robustness in some systems and metrics. In this paper, a thorough analysis of the results of systems submitted to recent editions of the IWSLT simultaneous track is provided to support our hypothesis and alternative ways to report latency metrics are proposed in order to provide a better understanding of SimulST systems' latency.

InImageTrans: Multimodal LLM-based Text Image Machine Translation

Fei Zuo, Kehai Chen, Yu Zhang, Zhengshan Xue, Min Zhang

Multimodal large language models (MLLMs) have shown remarkable capabilities across various downstream tasks. However, when MLLMs are transferred to the text image machine translation (TiMT) task, preliminary experiments reveal that MLLMs suffer from serious repetition and omission hallucinations. To alleviate these issues, this paper first designs an efficient MLLM named InImageTrans for TiMT and then proposes a simple and effective method named multi-conditional direct preference optimization (mcDPO) for advancing the TiMT. Particularly, the proposed mcDPO not only guides the MLLM in rejecting repetition output by creating text output preference pairs automatically, but also guides the MLLM in paying more attention to text information in images by creating image input preference pairs. Furthermore, we build a high-quality benchmark called MCiT for comprehensively evaluating the TiMT capabilities of InImageTrans. Experimental results show that the proposed method significantly outperforms existing open-source MLLMs on MCiT. All data and codes will be released once accepted.

LLMs Can Achieve High-quality Simultaneous Machine Translation as Efficiently as Offline

Biao Fu, Minpeng Liao, Kai Fan, Chengxi Li, Liang Zhang, Yidong Chen, Xiaodong Shi

When the complete source sentence is provided, Large Language Models (LLMs) perform excellently in offline machine translation even with a simple prompt "Translate the following sentence from [src lang] into [tgt lang]:". However, in many real scenarios, the source tokens arrive in a streaming manner and simultaneous machine translation (SiMT) is required, then the efficiency and performance of decoder-only LLMs are significantly limited by their auto-regressive nature. To enable LLMs to achieve high-quality SiMT as efficiently as offline translation, we propose a novel paradigm that includes constructing supervised fine-tuning (SFT) data for SiMT, along with new training and inference strategies. To replicate the token input/output stream in SiMT, the source and target tokens are rearranged into an interleaved sequence, separated by special tokens according to varying latency requirements. This enables powerful LLMs to learn read and write operations adaptively, based on varying latency prompts, while still maintaining efficient auto-regressive decoding. Experimental results show that, even with limited SFT data, our approach achieves state-of-the-art performance across various SiMT benchmarks and different evaluation metrics, and preserves the original capabilities of offline translation. Moreover, our approach generalizes well to document-level SiMT setting without requiring specific fine-tuning, even beyond the offline translation model.

Toward Global AI Inclusivity: A Large-Scale Multilingual Terminology Dataset (GIST)

Jiarui Liu, Iman Ouzzani, Wenkai Li, Lechen Zhang, Tianyue Ou, Houda Bouamor, Zhijing Jin, Mona T. Diab

The field of machine translation has achieved significant advancements, yet domain-specific terminology translation, particularly in AI, remains challenging. This work introduces GIST, a large-scale multilingual AI terminology dataset containing 5K terms extracted from top AI conference papers spanning 2000 to 2023. The terms were translated into Arabic, Chinese, French, Japanese, and Russian using a hybrid framework that combines LLMs for extraction with human expertise for translation. The dataset's quality was benchmarked against existing resources, demonstrating superior translation accuracy through crowdsourced evaluation. GIST was integrated into translation workflows using post-translation refinement methods that required no retraining, where LLM prompting consistently improved BLEU and COMET scores. A web demonstration on the ACL Anthology platform highlights its practical application, showcasing improved accessibility for non-English speakers. We address a critical gap in AI terminology resources and fosters global inclusivity and collaboration in AI research.

SignMusketeers: An Efficient Multi-Stream Approach for Sign Language Translation at Scale

Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, Karen Livescu

A persistent challenge in sign language video processing, including the task of sign language to written language translation, is how we train efficient model given the nature of videos. Informed by the nature and linguistics of signed languages, our proposed method focuses on just the most relevant parts in a signing video: the face, hands and body posture of the signer. However, instead of using pose estimation coordinates from off-the-shelf pose tracking models, which have inconsistent performance for hands and faces, we propose to learn the complex handshapes and rich facial expressions of sign languages in a self-supervised fashion. Our approach is based on learning from individual frames (rather than video sequences) and is therefore much more efficient than prior work on sign language pre-training. Compared to a recent model trained on publicly avaiable data that established a new state of the art in sign language translation on the How2Sign dataset, our approach yields similar translation performance, using less than 3% of the compute.

LegoMT2: Selective Asynchronous Sharded Data Parallel Training for Massive Neural Machine Translation

Fei Yuan, Yinquan Lu, Lei Li, Jingjing Xu

It is a critical challenge to learn a single model for massive languages. Prior methods focus on increasing the model size and training data size. However, large models are difficult to optimize efficiently even with distributed parallel training and translation capacity can interfere among languages. To address the challenge, we propose LegoMT2, an efficient training approach with an asymmetric multi-way model architecture for massive multilingual neural machine translation. LegoMT2 shards 435 languages into 8 language-centric groups and attributes one local encoder for each group's languages and a mix encoder-decoder for all languages. LegoMT2 trains the model through local data parallel and asynchronous distributed updating of parameters. LegoMT2 is 16.2\times faster than the distributed training method for M2M-100-12B (which only for 100 languages) while improving the translation results by an average of 2.2 BLEU on Flores-101. We will release the model and code to the public.

Improving MLLM's Document Image Machine Translation via Synchronously Self-reviewing Its OCR Proficiency

Yupu Liang, Yaping Zhang, Zhiyang Zhang, Zhiyuan Chen, Yang Zhao, Lu Xiang, Chengqing Zong, Yu Zhou

Multimodal Large Language Models (MLLMs) have shown strong performance in document image tasks, especially Optical Character Recognition (OCR). However, they struggle with Document Image Machine Translation (DIMT), which requires handling both cross-modal and cross-lingual challenges. Previous efforts to enhance DIMT capability through Supervised Fine-Tuning (SFT) on the DIMT dataset often result in the forgetting of the model's existing monolingual abilities, such as OCR. To address these challenges, we introduce a novel fine-tuning paradigm, named Synchronously Self-Reviewing (SSR) its OCR proficiency, inspired by the concept "Bilingual Cognitive Advantage". Specifically, SSR prompts the model to generate OCR text before producing translation text, which allows the model to leverage its strong monolingual OCR ability while learning to translate text across languages. Comprehensive experiments demonstrate the proposed SSR learning helps mitigate catastrophic forgetting, improving the generalization ability of MLLMs on both OCR and DIMT tasks. The code will be released upon acceptance.

Semantics-aware prompting for translating NOtices To AirMen

Minal Nitin Dani, Aishwarya Maheswaran, Maunendra Sankar Desarkar

A NOTAM or NOtice To AirMen is a crucial notice for different aviation stakeholders, particularly flight crews. It delivers essential notifications about abnormal conditions of Aviation System components such as changes to facilities, hazards, service, procedure that are not known far enough in advance to be publicized through other means. NOTAM messages are short, contain acronyms, and look cryptic in most of the cases. Writing and understanding these messages put heavy cognitive load on its end users. In this work, we take up the task of translating NOTAMs into English natural language using LLMs. Since NOTAMs do not adhere to English grammar rules and have their own decoding rules, large language models (LLMs) cannot translate them without effective prompting. In this paper, we develop a framework to come up with effective prompts to achieve the translations. Our approach uses context-aware semantic prompting techniques, paired with domain-specific rules, to improve the accuracy and clarity of translations. The framework is evaluated using comprehensive experiments (6 LLMs of varying sizes, and with 5 different prompting setups for each) and eight evaluation metrics measuring different aspects of the translation. The results demonstrate that our methodology can produce clear translations that accurately convey the information contained in NOTAMs.

BridG MT: Enhancing LLMs' Machine Translation Capabilities with Sentence Bridging and Gradual MT

Seungwoo Choi, Gahyun Yoo, Jay-Yoon Lee

Recent Large Language Models (LLMs) have demonstrated impressive translation performance without requiring fine-tuning on additional parallel corpora. However, they still face significant challenges in certain scenarios, particularly when translating low-resource languages. A common approach to address this issue is to provide external knowledge, such as few-shot examples, to assist LLMs in translating specific source sentences. However, this method is fundamentally limited by the quality or quantity of relevant sources, which cannot always be guaranteed. To reduce LLMs' reliance on external sources, we propose BridG MT, a method that combines Sentence Bridging, which generates a sequence of sentences as a bridge that gradually transition from easy-to-translate to more difficult, and Gradual MT, which sequentially translates these sentences using earlier translations as few-shot examples for subsequent ones. Experiments conducted on four LLMs across seven languages demonstrate that our method effectively enhances translation performance, achieving results that cannot be attained by simply providing a large number of few-shot examples.

A Measure of the System Dependence of Automated Metrics

Pius von Däniken, Jan Milan Deriu, Mark Cieliebak

Automated metrics for Machine Translation have made significant progress, with the goal of replacing expensive and time-consuming human evaluations. These metrics are typically assessed by their correlation with human judgments, which captures the monotonic relationship between human and metric scores. However, we argue that it is equally important to ensure that metrics treat all systems fairly and consistently. In this paper, we introduce a method to evaluate this aspect.

Did Translation Models Get More Robust Without Anyone Even Noticing?

Ben Peters, Andre Martins

Neural machine translation (MT) models achieve strong results across a variety of settings, but it is widely believed that they are highly sensitive to "noisy" inputs, such as spelling errors, abbreviations, and other formatting issues. In this paper, we revisit this insight in light of recent multilingual MT models and large language models (LLMs) applied to machine translation. Somewhat surprisingly, we show through controlled experiments that these models are far more robust to many kinds of noise than previous models, even when they perform similarly on clean data. This is notable because, even though LLMs have more parameters and more complex training processes than past models, none of the open ones we consider use any techniques specifically designed to encourage robustness. Next, we show that similar trends hold for social media translation experiments -- LLMs are more robust to social media text. We include an analysis of the circumstances in which source correction techniques can be used to mitigate the effects of noise. Altogether, we show that robustness to many types of noise has increased.

Unveiling the Power of Source: Source-based Minimum Bayes Risk Decoding for Neural Machine Translation

Boxuan Lyu, Hidetaka Kamigaito, Kotaro Funakoshi, Manabu Okumura

Maximum a posteriori decoding, a commonly used method for neural machine translation (NMT), aims to maximize the estimated posterior probability. However, high estimated probability does not always lead to high translation quality. Minimum Bayes Risk (MBR) decoding offers an alternative by seeking hypotheses with the highest expected utility. Inspired by Quality Estimation (QE) reranking which uses the QE model as a ranker, we propose source-based MBR (sMBR) decoding, a novel approach that utilizes quasi-sources (generated via paraphrasing or back-translation) as ``support hypotheses" and a reference-free quality estimation metric as the utility function, marking the first work to solely use sources in MBR decoding. Experiments show that sMBR outperforms QE reranking and the standard MBR decoding. Our findings suggest that sMBR is a promising approach for NMT decoding.

Fine-grained Video Dubbing Duration Alignment with Segment Supervised Preference Optimization

Chaoqun Cui, Liangbin Huang, Shijing Wang, Zhe Tong, Zhaolong Huang, Xiao Zeng, Xiaofeng Liu

Video dubbing aims to translate original speech in visual media programs from the source language to the target language, relying on neural machine translation and text-to-speech technologies. Due to varying information densities across languages, target speech often mismatches the source speech duration, causing audio-video synchronization issues that significantly impact viewer experience. In this study, we approach duration alignment in LLM-based video dubbing machine translation as a preference optimization problem. We propose the Segment Supervised Preference Optimization (SSPO) method, which employs a segment-wise sampling strategy and fine-grained loss to mitigate duration mismatches between source and target lines. Experimental results demonstrate that SSPO achieves superior performance in duration alignment tasks.

M-MAD: Multidimensional Multi-Agent Debate for Advanced Machine Translation Evaluation

Zhaopeng Feng, Jiayuan Su, Jiamei Zheng, Jiahan Ren, Yan Zhang, Jian Wu, Hongwei Wang, Zuozhu Liu

Recent advancements in large language models (LLMs) have given rise to the LLM-as-a-judge paradigm, showcasing their potential to deliver human-like judgments. However, in the field of machine translation (MT) evaluation, current LLM-as-a-judge methods fall short of learned automatic metrics. In this paper, we propose Multidimensional Multi-Agent Debate (M-MAD), a systematic LLM-based multi-agent framework for advanced LLM-as-a-judge MT evaluation. Our findings demonstrate that M-MAD achieves significant advancements by (1) decoupling heuristic MQM criteria into distinct evaluation dimensions for fine-grained assessments; (2) employing multi-agent debates to harness the collaborative reasoning capabilities of LLMs; (3) synthesizing dimension-specific results into a final evaluation judgment to ensure robust and reliable outcomes. Comprehensive experiments show that M-MAD not only outperforms all existing LLM-as-a-judge methods but also competes with state-of-the-art reference-based automatic metrics, even when powered by a suboptimal model like GPT-4o mini. Detailed ablations and analysis highlight the superiority of our framework design, offering a fresh perspective for LLM-as-a-judge paradigm. Our code and data are publicly available at https://github.com/SU-JIAYUAN/M-MAD.

Alleviating Distribution Shift in Synthetic Data for Machine Translation Quality Estimation

Xiang Geng, Zhejian Lai, Jiajun Chen, Hao Yang, Shujian Huang

Quality Estimation (QE) models evaluate the quality of machine translations without reference translations, serving as the reward models for the translation task. Due to the data scarcity, synthetic data generation has emerged as a promising solution. However, synthetic QE data often suffers from distribution shift, which can manifest as discrepancies between pseudo and real translations, or in pseudo labels that do not align with human preferences. To tackle this issue, we introduce DCSQE, a novel framework for alleviating distribution shift in synthetic QE data. To reduce the difference between pseudo and real translations, we employ the constrained beam search algorithm and enhance translation diversity through the use of distinct generation models. DCSQE uses references—i.e., translation supervision signals—to guide both the generation and annotation processes, enhancing the quality of token-level labels. DCSQE further identifies the shortest phrase covering consecutive error tokens, mimicking human annotation behavior, to assign the final phrase-level labels. Specially, we underscore that the translation model can not annotate translations of itself accurately. Extensive experiments demonstrate that DCSQE outperforms SOTA baselines like CometKiwi in both supervised and unsupervised settings. Further analysis offers insights into synthetic data generation that could benefit reward models for other tasks. The code is available at https://github.com/NJUNLP/njuqe.

Understanding In-context Machine Translation for Low-Resource Languages: A Case Study on Manchu

Renhao Pei, Yihong Liu, Peiqin Lin, François Yvon, Hinrich Schuetze

In-context machine translation (MT) with large language models (LLMs) is a promising approach for low-resource MT, as it can readily take advantage of linguistic resources such as grammar books and dictionaries. Such resources are usually selectively integrated into the prompt so that LLMs can directly perform translation without any specific training, via their in-context learning capability (ICL). However, the relative importance of each type of resource, e.g., dictionary, grammar book, and retrieved parallel examples, is not entirely clear. To address this gap, this study systematically investigates how each resource and its quality affect the translation performance, with the Manchu language as our case study. To remove any prior knowledge of Manchu encoded in the LLM parameters and single out the effect of ICL, we also experiment with an enciphered version of Manchu texts. Our results indicate that high-quality dictionaries and good parallel examples are very helpful, while grammars hardly help. In a follow-up study, we showcase a promising application of in-context MT: parallel data augmentation as a way to bootstrap a conventional MT model. When monolingual data abound, generating synthetic parallel data through in-context MT offers a pathway to mitigate data scarcity and build effective and efficient low-resource neural MT systems.

Locate-and-Focus: Enhancing Terminology Translation in Speech Language Models

Suhang Wu, Jialong Tang, Chengyi Yang, Pei Zhang, Baosong Yang, Junhui Li, Min Zhang, Jinsong Su

Direct speech translation (ST) has garnered increasing attention nowadays, yet the accurate translation of terminology within utterances remains a great challenge. In this regard, current studies mainly concentrate on leveraging various translation knowledge into ST models. However, these methods often struggle with interference from irrelevant noise and can not fully utilize the translation knowledge. To address these issues, in this paper, we propose a novel Locate-and-Focus method for terminology translation. It first effectively locates the speech clips containing terminologies within the utterance to construct translation knowledge, minimizing irrelevant information for the ST model. Subsequently, it associates the translation knowledge with the utterance and hypothesis from both audio and textual modalities, allowing the ST model to better focus on translation knowledge during translation. Experimental results across various datasets demonstrate that our method effectively locates terminologies within utterances and enhances the success rate of terminology translation, while maintaining robust general translation performance.

Single-to-mix Modality Alignment with Multimodal Large Language Model for Document Image Machine Translation

Yupu Liang, Yaping Zhang, Zhiyang Zhang, Yang Zhao, Lu Xiang, Chengqing Zong, Yu Zhou

Document Image Machine Translation (DIMT) aims to translate text within document images, facing generalization challenges due to limited training data and the complex interplay between visual and textual information. To address these challenges, we introduce M4Doc, a novel single-to-mix Modality alignment framework leveraging Multimodal Large Language Models (MLLMs). M4Doc aligns an imageonly encoder with the multimodal representations of an MLLM, pre-trained on large-scale document image datasets. This alignment enables a lightweight DIMT model to learn crucial visual-textual correlations during training. During inference, M4Doc bypasses the MLLM, maintaining computational efficiency while benefiting from its multimodal knowledge. Comprehensive experiments demonstrate substantial improvements in translation quality, especially in cross-domain generalization and challenging document image scenarios. The code will be released upon acceptance.

Making LLMs Better Many-to-Many Speech-to-Text Translators with Curriculum Learning

Yexing Du, Youcheng Pan, Ziyang Ma, Bo Yang, Yifan Yang, Keqi Deng, Xie Chen, Yang Xiang, Ming Liu, Bing Qin

Multimodal Large Language Models (MLLMs) have achieved significant success in Speech-to-Text Translation (S2TT) tasks. While most existing research has focused on English-centric translation directions, the exploration of many-to-many translation is still limited by the scarcity of parallel data. To address this, we propose a three-stage curriculum learning strategy that leverages the machine translation capabilities of large language models and adapts them to S2TT tasks, enabling effective learning in low-resource settings. We trained MLLMs with varying parameter sizes (3B, 7B, and 32B) and evaluated the proposed strategy using the FLEURS and CoVoST-2 datasets. Experimental results show that the proposed strategy achieves state-of-the-art average performance in 15\times14 language pairs, requiring fewer than 10 hours of speech data per language to achieve competitive results. The source code and models are released at .

Lost in Literalism: How Supervised Training Shapes Translationese in LLMs

Yafu Li, Ronghao Zhang, Zhilin Wang, Huajian Zhang, Leyang Cui, Yongjing Yin, Tong Xiao, Yue Zhang

Large language models (LLMs) have achieved remarkable success in machine translation, demonstrating impressive performance across diverse languages. However, translationese—characterized by overly literal and unnatural translations—remains a persistent challenge in LLM-based translation systems. Despite their pre-training on vast corpora of natural utterances, LLMs exhibit translationese errors and generate unexpected unnatural translations, stemming from biases introduced during supervised fine-tuning (SFT). In this work, we systematically evaluate the prevalence of translationese in LLM-generated translations and investigate its roots during supervised training. We introduce methods to mitigate these biases, including polishing golden references and filtering unnatural training instances. Empirical evaluations demonstrate that these approaches significantly reduce translationese while improving translation naturalness, validated by human evaluations and automatic metrics. Our findings highlight the need for training-aware adjustments to optimize LLM translation outputs, paving the way for more fluent and target-language-consistent translations.

Two Intermediate Translations Are Better Than One: Fine-tuning LLMs for Document-level Translation Refinement

Yichen Dong, Xinglin Lyu, Junhui Li, Daimeng Wei, Min Zhang, Shimin Tao, Hao Yang

Recent research has shown that large language models (LLMs) can enhance translation quality through self-refinement. In this paper, we build on this idea by extending the refinement from sentence-level to document-level translation, specifically focusing on document-to-document (Doc2Doc) translation refinement. Since sentence-to-sentence (Sent2Sent) and Doc2Doc translation address different aspects of the translation process, we propose fine-tuning LLMs for translation refinement using two intermediate translations, combining the strengths of both Sent2Sent and Doc2Doc. Additionally, recognizing that the quality of intermediate translations varies, we introduce an enhanced fine-tuning method with quality awareness that assigns lower weights to easier translations and higher weights to more difficult ones, enabling the model to focus on challenging translation cases. Experimental results across ten translation tasks with LLaMA-3-8B-Instruct and Mistral-Nemo-Instruct demonstrate the effectiveness of our approach. We will release our code on GitHub.

Multilingual Gloss-free Sign Language Translation: Towards Building a Sign Language Foundation Model

Sihan Tan, Taro Miyazaki, Kazuhiro Nakadai

Sign Language Translation (SLT) aims to convert sign language (SL) videos into spoken language text, thereby bridging the communication gap between the sign and the spoken community. While most existing works focus on translating a single SL into a single spoken language (one-to-one SLT), leveraging multilingual resources could mitigate low-resource issues and enhance accessibility. However, multilingual SLT (MLSLT) remains unexplored due to language conflicts and alignment difficulties across SLs and spoken languages. To address these challenges, we propose a multilingual gloss-free model with dual CTC objectives for token-level SL identification and spoken text generation. Our model supports 10 SLs and handles one-to-one, many-to-one, and many-to-many SLT tasks, achieving competitive performance compared to state-of-the-art methods on three widely adopted benchmarks: multilingual SP-10, PHOENIX14T, and CSL-Daily.

Theoretical Guarantees for Minimum Bayes Risk Decoding

Yuki Ichihara, Yuu Jinnai, Kaito Ariu, Tetsuro Morimura, Eiji Uchibe

Minimum Bayes Risk (MBR) decoding optimizes output selection by maximizing the expected utility value of an underlying human distribution. While prior work has shown the effectiveness of MBR decoding through empirical evaluation, few studies have analytically investigated why the method is effective. As a result of our analysis, we show that, given the size n of the reference hypothesis set used in computation, MBR decoding approaches the optimal solution with high probability at a rate of \mathcal{O}(n^{-\frac{1}{2}}), under certain assumptions, even though the language space \mathcal{Y} is significantly larger |\mathcal{Y}| \gg n. This result helps to theoretically explain the strong performance observed in several prior empirical studies on MBR decoding. In addition, we provide the performance gap for maximum-a-posteriori (MAP) decoding and compare it to MBR decoding. The result of this paper indicates that MBR decoding tends to converge to the optimal solution faster than MAP decoding in several cases.

DialUp! Modeling the Language Continuum by Adapting Models to Dialects and Dialects to Models

Niyati Bafna, Emily Chang, Nathaniel Romney Robinson, David R. Mortensen, Kenton Murray, David Yarowsky, Hale Sirin

Most of the world's languages and dialects are low-resource, and lack support in mainstream machine translation (MT) models. However, many of them have a closely-related high-resource language (HRL) neighbor, and differ in linguistically regular ways from it. This underscores the importance of model robustness to dialectal variation and cross-lingual generalization to the HRL dialect continuum. We present DialUp, consisting of a training-time technique for adapting a pretrained model to dialectal data (M-->D), and an inference-time intervention adapting dialectal data to the model expertise (D-->M). M-->D induces model robustness to potentially unseen and unknown dialects by exposure to synthetic data exemplifying linguistic mechanisms of dialectal variation, whereas D-->M treats dialectal divergence for known target dialects. These methods show considerable performance gains for several dialects from four language families, and modest gains for two other language families. We also conduct feature and error analyses, which show that language varieties with low baseline MT performance are more likely to benefit from these approaches.

THOR-MoE: Hierarchical Task-Guided and Context-Responsive Routing for Neural Machine Translation

Yunlong Liang, Fandong Meng, Jie Zhou

The sparse Mixture-of-Experts (MoE) has achieved significant progress for neural machine translation (NMT). However, there exist two limitations in current MoE solutions which may lead to sub-optimal performance: 1) they directly use the task knowledge of NMT into MoE (\emph{e.g.}, domain/linguistics-specific knowledge), which are generally unavailable at practical application and neglect the naturally grouped domain/linguistic properties; 2) the expert selection only depends on the localized token representation without considering the context, which fully grasps the state of each token in a global view. To address the above limitations, we propose THOR-MoE via arming the MoE with hierarchical task-guided and context-responsive routing policies. Specifically, it 1) firstly predicts the domain/language label and then extracts mixed domain/language representation to allocate task-level experts in a hierarchical manner; 2) injects the context information to enhance the token routing from the pre-selected task-level experts set, which can help each token to be accurately routed to more specialized and suitable experts. Extensive experiments on multi-domain translation and multilingual translation benchmarks with different architectures consistently demonstrate the superior performance of THOR-MoE. Additionally, the THOR-MoE operates as a plug-and-play module compatible with existing Top-\cite{shazeer2017} or Top-\cite{huang-etal-2024-harder} routing schemes, ensuring broad applicability across diverse MoE architectures. For instance, compared with vanilla Top-\cite{huang-etal-2024-harder} routing, the context-aware manner can achieve an average improvement of 0.75 BLEU with less than 22% activated parameters on multi-domain translation tasks.

Enhancing Machine Translation with Self-Supervised Preference Data

Haoxiang Sun, Ruize Gao, Pei Zhang, Baosong Yang, Rui Wang

Model alignment methods like Direct Preference Optimization and Contrastive Preference Optimization have enhanced machine translation performance by leveraging preference data to enable models to reject suboptimal outputs. During preference data construction, previous approaches primarily rely on humans, strong models like GPT4 or model self-sampling. In this study, we first explain the shortcomings of this practice. Then, we propose Self-Supervised Preference Optimization (SSPO), a novel framework which efficiently constructs translation preference data for iterative DPO training. Applying SSPO to 14B parameters large language models (LLMs) achieves comparable or better performance than GPT-4o on FLORES and multi-domain test datasets. We release an augmented MQM dataset in https://github.com/sunny-sjtu/MQM-aug.

Multi-perspective Alignment for Increasing Naturalness in Neural Machine Translation

Huiyuan Lai, Esther Ploeger, Rik van Noord, Antonio Toral

Neural machine translation (NMT) systems amplify lexical biases present in their training data, leading to artificially impoverished language in output translations. These language-level characteristics render automatic translations different from text originally written in a language and human translations, which hinders their usefulness in for example creating evaluation datasets. Attempts to increase naturalness in NMT can fall short in terms of content preservation, where increased lexical diversity comes at the cost of translation accuracy. Inspired by the reinforcement learning from human feedback framework, we introduce a novel method that rewards both naturalness and content preservation. We experiment with multiple perspectives to produce more natural translations, aiming at reducing machine and human translationese. We evaluate our method on English-to-Dutch literary translation, and find that our best model produces translations that are lexically richer and exhibit more properties of human-written language, without loss in translation accuracy.

Document-Level Text Generation with Minimum Bayes Risk Decoding using Optimal Transport

Yuu Jinnai

Document-level text generation tasks are known to be more difficult than sentence-level text generation tasks as they require an understanding of longer context to generate high-quality texts. In this paper, we investigate the adaptation of Minimum Bayes Risk (MBR) decoding for document-level text generation tasks. MBR decoding makes use of a utility function to estimate the output with the highest expected utility from a set of candidate outputs. Although MBR decoding is shown to be effective in a wide range of sentence-level text generation tasks, its performance on document-level text generation tasks is limited, as many of the utility functions are designed for evaluating the utility of sentences. To this end, we propose MBR-OT, a variant of MBR decoding using Wasserstein distance to compute the utility of a document using a sentence-level utility function. The experimental result shows that the performance of MBR-OT outperforms that of the standard MBR in document-level machine translation, text simplification, and dense image captioning tasks.

Comprehensive Analysis of Minimum Bayes Risk Decoding through Bias and Diversity Decomposition

Hidetaka Kamigaito, Hiroyuki Deguchi, Yusuke Sakai, Katsuhiko Hayashi, Taro Watanabe

Inference methods play an important role in eliciting the performance of large language models (LLMs). Currently, LLMs use inference methods utilizing generated multiple samples, which can be derived from Minimum Bayes Risk (MBR) Decoding. Previous studies have conducted empirical analyses to clarify the improvements in generation performance achieved by MBR decoding and have reported various observations. However, the theoretical underpinnings of these findings remain uncertain. To address this, we offer a new theoretical interpretation of MBR decoding from the perspective of bias–diversity decomposition. In this interpretation, the error in the quality estimation of hypotheses by MBR decoding is decomposed into two main factors: bias, which considers the closeness between the utility function and human evaluation, and diversity, which represents the variability in the quality estimation of the utility function. The theoretical analysis reveals the difficulty of simultaneously improving bias and diversity, confirming the validity of enhancing MBR decoding performance by increasing diversity. Furthermore, we reveal that diversity can explain one aspect of inference scaling laws that describe performance improvement by increasing sample size. Moreover, experiments across multiple NLP tasks yielded results consistent with these theoretical characteristics. Our code is available at https://github.com/naist-nlp/mbr-bias-diversity.

An Empirical Study of Iterative Refinements for Non-autoregressive Translation

Yisheng Xiao, Pei Guo, Zechen Sun, Juntao Li, Kai Song, Min Zhang

Iterative non-autoregressive (NAR) models share a spirit of mixed autoregressive (AR) and fully NAR models, seeking a balance between generation quality and inference efficiency. These models have recently demonstrated impressive performance in varied generation tasks, surpassing the autoregressive Transformer. However, they also face several challenges that impede further development. In this work, we target building more efficient and competitive iterative NAR models. Firstly, we produce two simple metrics to identify the potential problems existing in current refinement processes, and look back on the various iterative NAR models to find the key factors for realizing our purpose. Subsequently, based on the analyses of the limitations of previous inference algorithms, we propose a simple yet effective strategy to conduct efficient refinements without performance declines. Experiments on five widely used datasets show that our final models set the new state-of-the-art performance compared to all previous NAR models, even with fewer decoding steps, and outperform AR Transformer by around one BLEU on average. Our codes and models are available on Github.

GrammaMT: Improving Machine Translation with Grammar-Informed In-Context Learning

Rita Ramos, Everlyn Asiko Chimoto, Maartje Ter Hoeve, Natalie Schluter

We introduce GrammaMT, a grammatically-aware prompting approach for machine translation that uses Interlinear Glossed Text (IGT), a common form of linguistic description providing morphological and lexical annotations for source sentences. GrammaMT proposes three prompting strategies: gloss-shot, chain-gloss and model-gloss. All are training-free, requiring only a few examples that involve minimal effort to collect, and making them well-suited for low-resource setups. Experiments show that GrammaMT enhances translation performance on open-source instruction-tuned LLMs for various low- to high-resource languages across three benchmarks: (1) the largest IGT corpus, (2) the challenging 2023 SIGMORPHON Shared Task data over endangered languages, and (3) even in an out-of-domain setting with FLORES. Moreover, ablation studies reveal that leveraging gloss resources could substantially boost MT performance (by over 17 BLEU points) if LLMs accurately generate or access input sentence glosses.

Session 10: Machine Translation (Tuesday 29 July, 16:00-17:30, Hall 5X)

Tulun: Transparent and Adaptable Low-resource Machine Translation

Raphael Merx, Hanna Suominen, Lois Yinghui HONG, Nick Thieberger, Trevor Cohn, Ekaterina Vylomova

Machine translation (MT) systems that support low-resource languages often struggle on specialized domains. While researchers have proposed various techniques for domain adaptation, these approaches typically require model fine-tuning, making them impractical for non-technical users and small organizations. To address this gap, we propose Tulun, a versatile solution for terminology-aware translation, combining neural MT with large language model (LLM)-based post-editing guided by existing glossaries and translation memories. Our open-source web-based platform enables users to easily create, edit, and leverage terminology resources, fostering a collaborative human-machine translation process that respects and incorporates domain expertise while increasing MT accuracy. Evaluations show effectiveness in both real-world and benchmark scenarios: on medical and disaster relief translation tasks for Tetun and Bislama, our system achieves improvements of 16.90-22.41 ChrF++ points over baseline MT systems. Across six low-resource languages on the FLORES dataset, Tulun outperforms both standalone MT and LLM approaches, achieving an average improvement of 2.8 ChrF points over NLLB-54B. Tulun is publicly accessible at https://bislama-trans.rapha.dev.

Session 10: Multimodality and Language Grounding to Vision, Robotics and Beyond (Tuesday 29 July, 16:00-17:30, Hall 4/5)

Vulnerability of Text-to-Image Models to Prompt Template Stealing: A Differential Evolution Approach

Yurong Wu, Fangwen Mu, Qiuhong Zhang, Jinjing Zhao, Xinrun Xu, Lingrui Mei, Yang Wu, Lin Shi, Junjie Wang, Zhiming Ding, Yiwei Wang

Prompt trading has emerged as a significant intellectual property concern in recent years, where vendors entice users by showcasing sample images before selling prompt templates that can generate similar images. This work investigates a critical security vulnerability: attackers can steal prompt templates using only a limited number of sample images. To investigate this threat, we introduce \textbf{Prism}, a prompt-stealing benchmark consisting of 50 templates and 450 images, organized into Easy and Hard difficulty levels. To identify the vulnerabity of VLMs to prompt stealing, we propose \textbf{EvoStealer}, a novel template stealing method that operates without model fine-tuning by leveraging differential evolution algorithms. The system first initializes population sets using multimodal large language models (MLLMs) based on predefined patterns, then iteratively generates enhanced offspring through MLLMs. During evolution, EvoStealer identifies common features across offspring to derive generalized templates. Our comprehensive evaluation conducted across open-source (InternVL2-26B) and closed-source models (GPT-4o and GPT-4o-mini) demonstrates that EvoStealer's stolen templates can reproduce images highly similar to originals and effectively generalize to other subjects, significantly outperforming baseline methods with an average improvement of over 10%. Moreover, our cost analysis reveals that EvoStealer achieves template stealing with negligible computational expenses. Our code and dataset are available at https://whitepagewu.github.io/evostealer-site.

Change Entity-guided Heterogeneous Representation Disentangling for Change Captioning

Yi Li, Yunbin Tu, Liang Li, Li Su, Qingming Huang

Change captioning aims to describe differences between a pair of images using natural language. However, learning effective difference representations is highly challenging due to distractors such as illumination and viewpoint changes. To address this, we propose a change-entity-guided disentanglement network that explicitly learns difference representations while mitigating the impact of distractors. Specifically, we first design a change entity retrieval module to identify key objects involved in the change from a textual perspective. Then, we introduce a difference representation enhancement module that strengthens the learned features, disentangling genuine differences from background variations. To further refine the generation process, we incorporate a gated Transformer decoder, which dynamically integrates both visual difference and textual change-entity information. Extensive experiments on CLEVR-Change, CLEVR-DC and Spot-the-Diff datasets demonstrate that our method outperforms existing approaches, achieving state-of-the-art performance. The code will be released.

VP-MEL: Visual Prompts Guided Multimodal Entity Linking

Hongze Mi, Jinyuan Li, Zhangxuying, Haoran Cheng, Jiahao Wang, Di Sun, Gang Pan

Multimodal entity linking (MEL), a task aimed at linking mentions within multimodal contexts to their corresponding entities in a knowledge base (KB), has attracted much attention due to its wide applications in recent years. However, existing MEL methods often rely on mention words as retrieval cues, which limits their ability to effectively utilize information from both images and text. This reliance causes MEL to struggle with accurately retrieving entities in certain scenarios, especially when the focus is on image objects or mention words are missing from the text. To solve these issues, we introduce a Visual Prompts guided Multimodal Entity Linking (VP-MEL) task. Given a text-image pair, VP-MEL aims to link a marked region (i.e., visual prompt) in an image to its corresponding entities in the knowledge base. To facilitate this task, we present a new dataset, VPWiki, specifically designed for VP-MEL. Furthermore, we propose a framework named IIER, which enhances visual feature extraction using visual prompts and leverages the pre-trained Detective-VLM model to capture latent information. Experimental results on the VPWiki dataset demonstrate that IIER outperforms baseline methods across multiple benchmarks for the VP-MEL task.

Libra: Leveraging Temporal Images for Biomedical Radiology Analysis

Xi Zhang, Zaiqiao Meng, Jake Lever, Edmond S. L. Ho

Radiology report generation (RRG) requires advanced medical image analysis, effective temporal reasoning, and accurate text generation. While multimodal large language models (MLLMs) align with pre-trained vision encoders to enhance visual-language understanding, most existing methods rely on single-image analysis or rule-based heuristics to process multiple images, failing to fully leverage temporal information in multi-modal medical datasets. In this paper, we introduce **Libra**, a temporal-aware MLLM tailored for chest X-ray report generation. Libra combines a radiology-specific image encoder with a novel Temporal Alignment Connector (**TAC**), designed to accurately capture and integrate temporal differences between paired current and prior images. Extensive experiments on the MIMIC-CXR dataset demonstrate that Libra establishes a new state-of-the-art benchmark among similarly scaled MLLMs, setting new standards in both clinical relevance and lexical accuracy. All source code and data are publicly available at: https://github.com/X-iZhang/Libra.

Look \& Mark: Leveraging Radiologist Eye Fixations and Bounding boxes in Multimodal Large Language Models for Chest X-ray Report Generation

Yunsoo Kim, Jinge Wu, Su Hwan Kim, Pardeep Vasudev, Jiashu Shen, Honghan Wu

Recent advancements in multimodal Large Language Models (LLMs) have significantly enhanced the automation of medical image analysis, particularly in generating radiology reports from chest X-rays (CXR). However, these models still suffer from hallucinations and clinically significant errors, limiting their reliability in real-world applications. In this study, we propose Look \& Mark (L\&M), a novel grounding fixation strategy that integrates radiologist eye fixations (Look) and bounding box annotations (Mark) into the LLM prompting framework. Unlike conventional fine-tuning, L\&M leverages in-context learning to achieve substantial performance gains without retraining. When evaluated across multiple domain-specific and general-purpose models, L\&M demonstrates significant gains, including a 1.2% improvement in overall metrics (A.AVG) for CXR-LLaVA compared to baseline prompting and a remarkable 9.2% boost for LLaVA-Med. General-purpose models also benefit from L\&M combined with in-context learning, with LLaVA-OV achieving an 87.3% clinical average performance (C.AVG)—the highest among all models, even surpassing those explicitly trained for CXR report generation. Expert evaluations further confirm that L\&M reduces clinically significant errors (by 0.43 average errors per report), such as false predictions and omissions, enhancing both accuracy and reliability. These findings highlight L\&M's potential as a scalable and efficient solution for AI-assisted radiology, paving the way for improved diagnostic workflows in low-resource clinical settings.

Learning to Better Act by Post-training on Vision Language Tasks

Li Muyao, Zihao Wang, Kaichen He, Xiaojian Ma, Yitao Liang

Recently, action-based decision-making in open-world environments has gained significant attention. Visual Language Action (VLA) models, pretrained on large-scale web datasets, have shown promise in decision-making tasks. However, previous work has primarily focused on action post-training, often neglecting enhancements to the foundation model itself. In response, we introduce a novel approach, **Act from Visual Language Post-Training**, which refines Visual Language Models (VLMs) through visual and linguistic guidance in a self-supervised manner. This enhancement improves the models' capabilities in world knowledge, visual recognition, and spatial grounding in open-world environments. Our experiments demonstrate that post-training on non-trajectory tasks leads to a significant 40% improvement over the best agent baseline on a diverse set of atomic tasks. Additionally, we show that our approach outperforms traditional imitation learning-based VLA models in Minecraft. We will open-source the code, models, and datasets to foster further research.

Multimodal Inconsistency Reasoning (MMIR): A New Benchmark for Multimodal Reasoning Models

Qianqi Yan, Yue Fan, Hongquan Li, Shan Jiang, Yang Zhao, Xinze Guan, Ching-Chen Kuo, Xin Eric Wang
Existing Multimodal Large Language Models (MLLMs) are predominantly trained and tested on consistent visual-textual inputs, leaving open the question of whether they can handle inconsistencies in real-world, layout-rich content. To bridge this gap, we propose the Multimodal Inconsistency Reasoning (MMIR) benchmark to assess MLLMs' ability to detect and reason about semantic mismatches in artifacts such as webpages, presentation slides, and posters. MMIR comprises 534 challenging samples, each containing synthetically injected errors across five reasoning-heavy categories: Factual Contradiction, Identity Misattribution, Contextual Mismatch, Quantitative Discrepancy, and Temporal/Spatial Incoherence. We evaluate six state-of-the-art MLLMs, showing that models with dedicated multimodal reasoning capabilities, such as o1, substantially outperform their counterparts while open-source models remain particularly vulnerable to inconsistency errors. Detailed error analyses further show that models excel in detecting inconsistencies confined to a single modality, particularly in text, but struggle with cross-modal conflicts and complex layouts. Probing experiments reveal that single-modality prompting, including Chain-of-Thought (CoT) and Set-of-Mark (SoM) methods, yields marginal gains, revealing a key bottleneck in cross-modal reasoning. Our findings highlight the need for advanced multimodal reasoning and point to future research on multimodal inconsistency.

Vision-Language Models Struggle to Align Entities across Modalities

Iñigo Alonso, Gorka Azkune, Ander Salaberria, Jeremy Barnes, Oier Lopez de Lacalle
Cross-modal entity linking refers to the ability to align entities and their attributes across different modalities. While cross-modal entity linking is a fundamental skill needed for real-world applications such as multimodal code generation, fake news detection, or scene understanding, it has not been thoroughly studied in the literature. In this paper, we introduce a new task and benchmark to address this gap. Our benchmark, MATE, consists of 5.5k evaluation instances featuring visual scenes aligned with their textual representations. To evaluate cross-modal entity linking performance, we design a question-answering task that involves retrieving one attribute of an object in one modality based on a unique attribute of that object in another modality. We evaluate state-of-the-art Vision-Language Models (VLMs) and humans on this task, and find that VLMs struggle significantly compared to humans, particularly as the number of objects in the scene increases. Our analysis also shows that, while chain-of-thought prompting can improve VLM performance, models remain far from achieving human-level proficiency. These findings highlight the need for further research in cross-modal entity linking and show that MATE is a strong benchmark to support that progress.

ChartQAPro: A More Diverse and Challenging Benchmark for Chart Question Answering

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, Megh Thakkar, Md Rizwan Parvez, Enamul Hoque, Shafiq Joty
Charts are ubiquitous, as people often use them to analyze data, answer questions, and discover critical insights. However, performing complex analytical tasks with charts requires significant perceptual and cognitive effort. Chart Question Answering (CQA) systems automate this process by enabling models to interpret and reason with visual representations of data. However, existing benchmarks like ChartQA lack real-world diversity and have recently shown performance saturation with modern large vision-language models (LVLMs). To address these limitations, we introduce ChartQAPro, a new benchmark that includes 1,341 charts from 157 diverse sources, spanning various chart types—including infographics and dashboards—and featuring 1,948 questions in various types, such as multiple-choice, conversational, hypothetical, and unanswerable questions, to better reflect real-world challenges. Our evaluations with 21 models show a substantial performance drop for LVLMs on ChartQAPro; e.g., Claude Sonnet 3.5 scores 90.5% on ChartQA but only 55.81% on ChartQAPro, underscoring the complexity of chart reasoning. We complement our findings with detailed error analyses and ablation studies, identifying key challenges and opportunities for advancing LVLMs in chart understanding and reasoning. We release ChartQAPro at <redacted>.

Worse than Random? An Embarrassingly Simple Probing Evaluation of Large Multimodal Models in Medical VQA

Qianqi Yan, Xuehai He, Xiang Yue, Xin Eric Wang

Large Multimodal Models (LMMs) have demonstrated impressive performance on existing medical Visual Question Answering (Med-VQA) benchmarks. However, high reported accuracy does not necessarily reflect their true diagnostic reliability in clinical settings. This study reveals that state-of-the-art models perform worse than random guessing on medical diagnosis questions when subjected to simple Probing Evaluation for Medical Diagnosis (ProbMed). ProbMed challenges models through probing evaluation and procedural diagnosis. Particularly, probing evaluation features pairing ground-truth questions with adversarial counterparts that feature negated and hallucinated attributes, while procedural diagnosis requires reasoning across various dimensions for each image, including modality recognition, organ identification, clinical findings, abnormalities, and positional grounding. Our evaluation reveals that even top-performing models like GPT-4o, GPT-4V, and Gemini Pro perform worse than random guessing on specialized diagnostic questions, indicating significant limitations in handling fine-grained medical inquiries. Furthermore, our ablation study on open-source models (e.g., LLaVA, LLaVA-Med, and Med-Flamingo) identifies poor visual understanding as a primary bottleneck—a limitation that can be partially mitigated by incorporating visual descriptions generated by GPT-4o, resulting in an average performance improvement of 9.44%. These findings underscore the urgent need for more robust evaluation methods and domain-specific expertise to ensure the reliability of LMMs in high-stakes medical applications.

EgoNormia: Benchmarking Physical Social Norm Understanding

MohammadHossein Rezaei, Yicheng Fu, Phil Cuvin, Caleb Ziems, Yanzhe Zhang, Hao Zhu, Diyi Yang

Human activity is moderated by norms. However, machines are often trained without explicit supervision on norm understanding and reasoning, especially when the norms are grounded in a physical and social context. To improve and evaluate the normative reasoning capability of vision-language models (VLMs), we present EgoNormia, consisting of 1,853 ego-centric videos of human interactions, each of which has two related questions evaluating both the prediction and justification of normative actions. The normative actions encompass seven categories: safety, privacy, proxemics, politeness, cooperation, coordination/proactivity, and communication/legibility. To compile this dataset at scale, we propose a novel pipeline leveraging video sampling, automatic answer generation, filtering, and human validation. Our work demonstrates that current state-of-the-art vision-language models lack robust norm understanding, scoring a maximum of 54% on EgoNormia (versus a human bench of 92%). Our analysis of performance in each dimension highlights the significant risks of safety, privacy, and the lack of collaboration and communication capability when applied to real-world agents. We additionally show that through a retrieval-based generation method, it is possible to use EgoNormia to enhance normative reasoning in VLMs.

CRAB: Cross-environment Agent Benchmark for Multimodal Language Model Agents

Tianqi Xu, Linyao Chen, Dai-Jie Wu, Yanjun Chen, Zecheng Zhang, Xiang Yao, Zhiqiang Xie, Yongchao Chen, Shilong Liu, Bochen Qian, Anjie Yang, Zhaoxuan Jin, Jianbo Deng, Philip Torr, Bernard Ghanem, Guohao Li

The development of autonomous agents increasingly relies on Multimodal Language Models (MLMs) to perform tasks described in natural language with GUI environments, such as websites, desktop computers, or mobile phones. Existing benchmarks for MLM agents in interactive environments are limited by their focus on a single environment, lack of detailed and generalized evaluation methods, and the complexities of constructing tasks and evaluators. To overcome these limitations, we introduce CRAB, the first cross-environment agent benchmark framework, incorporating a graph-based fine-grained evaluation method and an efficient task generation method. Our framework supports multiple devices and can be easily extended to any environment with a Python interface. Leveraging CRAB, we develope CRAB Benchmark-v0 comprising 120 tasks in computer desktop and mobile phone environments. We evaluated 6 advanced MLMs using different single and multi-agent system configurations on this benchmark. The experimental results demonstrate that the single agent with GPT-4o achieves the best completion ratio of 38.01%.

KITAB-Bench: A Comprehensive Multi-Domain Benchmark for Arabic OCR and Document Understanding

Ahmed Heakl, Muhammad Abdullah Sohail, Mukul Ranjan, Rania Elbadry, Ghazi Shazan Ahmad, Mohamed El-Geish, Omar Maher, Zhiqiang Shen, Fahad Shahbaz Khan, Salman Khan

With the growing adoption of Retrieval-Augmented Generation (RAG) in document processing, robust text recognition has become increasingly critical for knowledge extraction. While OCR (Optical Character Recognition) for English and other languages benefits from large datasets and well-established benchmarks, Arabic OCR faces unique challenges due to its cursive script, right-to-left text flow, and complex typographic and calligraphic features. We present KITAB-Bench, a comprehensive Arabic OCR benchmark that fills the gaps in current evaluation systems. Our benchmark comprises 8,809 samples across 9 major domains and 36 sub-domains, encompassing diverse document types including handwritten text, structured tables, and specialized coverage of 21 chart types for business intelligence. Our findings show that modern vision-language models (such as GPT-4, Gemini, and Qwen) outperform traditional OCR approached (like EasyOCR, PaddleOCR, and Surya) by an average of 60% in Character Error Rate (CER). Furthermore, we highlight significant limitations of current Arabic OCR models, particularly in PDF-to-Markdown conversion, where the best model Gemini-2.0-Flash achieves only 65% accuracy. This underscores the challenges in accurately recognizing Arabic text, including issues with complex fonts, numeral recognition errors, word elongation, and table structure detection. This work establishes a rigorous evaluation framework that can drive improvements in Arabic document analysis methods and bridge the performance gap with English OCR technologies.

BottleHumor: Self-Informed Humor Explanation using the Information Bottleneck Principle

EunJeong Hwang, Peter West, Vered Shwartz

Humor is prevalent in online communications and it often relies on more than one modality (e.g., cartoons and memes). Interpreting humor in multimodal settings requires drawing on diverse types of knowledge, including metaphorical, sociocultural, and commonsense knowledge. However, identifying the most useful knowledge remains an open question. We introduce BottleHumor, a method inspired by the information bottleneck principle that elicits relevant world knowledge from vision and language models which is iteratively refined for generating an explanation of the humor in an unsupervised manner. Our experiments on three datasets confirm the advantage of our method over a range of baselines. Our method can further be adapted in the future for additional tasks that can benefit from eliciting and conditioning on relevant world knowledge and open new research avenues in this direction.

LADDER: Language-Driven Slice Discovery and Error Rectification in Vision Classifiers

Shantanu Ghosh, Rayan Syed, Chenyu Wang, Vaibhav Choudhary, Binxu Li, Clare B Poynton, Shyam Visweswaran, kayhan Batmanghelich

Slice discovery refers to identifying systematic biases in the mistakes of pre-trained vision models. Current slice discovery methods in computer vision rely on converting input images into sets of attributes and then testing hypotheses about configurations of these pre-computed attributes associated with elevated error patterns. However, such methods face several limitations: 1) they are restricted by the predefined attribute bank; 2) they lack the common sense reasoning and domain-specific knowledge often required for specialized fields e.g., radiology; 3) at best, they can only identify biases in image attributes while overlooking those introduced during preprocessing or data preparation. We hypothesize that bias-inducing variables leave traces in the form of language (e.g., logs), which can be captured as unstructured text. Thus, we introduce ladder, which leverages the reasoning capabilities and latent domain knowledge of Large Language Models (LLMs) to generate hypotheses about these mistakes. Specifically, we project the internal activations of a pre-trained model into text using a retrieval approach and prompt the LLM to propose potential bias hypotheses. To detect biases from preprocessing pipelines, we convert the preprocessing data into text and prompt the LLM. Finally, ladder generates pseudo-labels for each identified bias, thereby mitigating all biases without requiring expensive attribute annotations. Rigorous evaluations on 3 natural and 3 medical imaging datasets, 200+ classifiers, and 4 LLMs with varied architectures and pretraining strategies - demonstrate that\ladder consistently outperforms current methods. Code is available: .

Seeing What Tastes Good: Revisiting Multimodal Distributional Semantics in the Billion Parameter Era

Dan Oneata, Desmond Elliott, Stella Frank

There is widespread agreement about the grounded nature of human learning and representation, and the belief that computational models of meaning need to be multimodal. In this paper, we ask to what degree does this belief hold in the era of models trained on billions of examples? We investigate the ability of pre-trained vision models to represent the semantic feature norms of concrete object concepts, e.g. a ROSE is red, smells sweet, and is a flower. More specifically, we use probing tasks to test which properties of objects these models are aware of. We evaluate image encoders trained on image data alone, as well as multimodally-trained image encoders and language-only models, on predicting an extended set of the classic McRae norms and the newer Binder dataset of attribute ratings. We find that multimodal image encoders slightly outperform language-only approaches, and that image-only encoders perform comparably to the language models, even on non-visual attributes that are classified as “encyclopedic” or “function”. These results offer new insights into what can be learned from pure unimodal learning, and the complementarity of the modalities.

Burn After Reading: Do Multimodal Large Language Models Truly Capture Order of Events in Image Sequences?

Yingjin Song, Yupei Du, Denis Paperno, Albert Gatt

This paper introduces the new and challenging TempVS benchmark, which focuses on temporal grounding and reasoning capabilities of Multimodal Large Language Models (MLLMs) in image sequences. TempVS consists of three main tests (i.e., event relation inference, sentence ordering and image ordering) and each accompanied with a basic grounding test, yielding a total of 2,085 annotated image sequences and 15k multiple-choice questions. TempVS requires MLLMs to rely on both visual and linguistic modalities to understand the temporal order of events. We extensively evaluate 38 state-of-the-art MLLMs, demonstrating that models struggle to solve TempVS. Our analysis reveals a substantial performance gap between current MLLMs and human capabilities, accompanied by fine-grained insights that suggest promising directions for future research.

Can VLMs Actually See and Read? A Survey on Modality Collapse in Vision-Language Models

Mong Yuan Sim, Wei Emma Zhang, Xiang Dai, Biaoyan Fang

Vision-language models (VLMs) integrate textual and visual information, enabling the model to process visual inputs and leverage visual information to generate predictions. Such models are demanding for tasks such as visual question answering, image captioning and visual grounding. However, some recent work found that %visual information is not fully utilized by language models. That is, VLMs often rely heavily on textual information, ignoring visual information, but are still able to achieve competitive performance in vision-language (VL) tasks. This survey reviews modality collapse analysis work to provide insights into the reason for this unintended behaviour. It also reviews probing studies for fine-grained vision-language understanding, presenting current findings on information encoded in VL representations and highlight potential directions for future research.

From Perception to Reasoning: Enhancing Vision-Language Models for Mobile UI Understanding

Settaluri Lakshmi Sravanthi, Ankit Mishra, Debjyoti Mondal, Subhadarshi Panda, Rituraj Singh, Pushpak Bhattacharyya

Accurately grounding visual and textual elements within mobile user interfaces (UIs) remains a significant challenge for Vision-Language Models (VLMs). Visual grounding, a critical task in this domain, involves identifying the most relevant UI element or region based on a natural language query—a process that requires both precise perception and context-aware reasoning. In this work, we present - **MoUI**, a light-weight mobile UI understanding model trained on **MoIT**, an instruction-tuning dataset specifically tailored for mobile screen understanding and grounding, designed to bridge the gap between user intent and visual semantics. Complementing this dataset, we also present a human-annotated reasoning benchmark **MoIQ** that rigorously evaluates complex inference capabilities over mobile UIs. To harness these resources effectively, we propose a two-stage training approach that separately addresses perception and reasoning tasks, leading to stronger perception capabilities and improvement in reasoning abilities. Through extensive experiments, we demonstrate that our MoUI models achieve significant gains in accuracy across all perception tasks and _state-of-the-art_ results on public reasoning benchmark **ComplexQA (78%) and our MoIQ (49%)**. We will be open-sourcing our dataset, code, and models to foster further research and innovation in the field.

EXPERT: An Explainable Image Captioning Evaluation Metric with Structured Explanations

Hyunjong Kim, Sangyeop Kim, Jongheon Jeong, Yeongjae Cho, Sungzoon Cho

Recent explainable metrics for image captioning aim to provide both scores and explanations for their evaluations. However, these metrics generate explanations without standardized criteria, and the overall quality of the generated explanations remains unverified. In this paper, we propose EXPERT, a reference-free evaluation metric that provides structured explanations based on three fundamental criteria: fluency, relevance, and descriptiveness. By constructing large-scale datasets of high-quality structured explanations, we develop a two-stage evaluation template to effectively supervise a vision-language model for both scoring and explanation generation. EXPERT achieves state-of-the-art results on benchmark datasets while providing significantly higher-quality explanations than existing metrics, as validated through comprehensive human evaluation.

Can Vision Language Models Understand Mimes?

Hyundong Justin Cho, Spencer Lin, Tejas Srinivasan, Michael Saxon, Deuksin Kwon, Natali T. Chavez, Jonathan May

Non-verbal communication (NVC) is an integral part of human language, but it has been overlooked in natural language processing research. Studying NVC in general is challenging because of its high variance in interpretation among individuals and cultures, but mime---the theatrical technique of suggesting intent using only gesture, expression, and movement---is a subset of NVC with much lower human interpretation variance. As a gateway for evaluating vision-language models on their understanding of NVC, we propose Mime Identification-based Multimodal Evaluation (MIME), a gesture recognition task built upon a novel corpus of mimed activity comprising 86 unique gestures with a variety of perturbations applied to the avatar, background, and viewpoint for evaluating recognition robustness. We find that both open-weight and API-based vision-language models perform significantly worse than humans at identifying mimed gestures in MIME, motivating the need for increased research for instilling more robust understanding of human actions for VLMs.

The Impact of Auxiliary Patient Data on Automated Chest X-Ray Report Generation and How to Incorporate It

Aaron Nicolson, Shengyao Zhuang, Jason Dowling, Bevan Koopman

This study investigates the integration of diverse patient data sources into multimodal language models for automated chest X-ray (CXR) report generation. Traditionally, CXR report generation relies solely on data from a patient's CXR exam, overlooking valuable information from patient electronic health records. Utilising the MIMIC-CXR and MIMIC-IV-ED datasets, we investigate the use of patient data from emergency department (ED) records — such as vital signs measured and medicines reconciled during an ED stay — for CXR report generation, with the aim of enhancing diagnostic accuracy. We also investigate conditioning CXR report generation on the clinical history section of radiology reports, which has been overlooked in the literature. We introduce a novel approach to transform these heterogeneous data sources into patient data embeddings that prompt a multimodal language model (CXRMate-ED). Our comprehensive evaluation indicates that using a broader set of patient data significantly enhances diagnostic accuracy. The model, training code, and dataset are publicly available.

Synergizing LLMs with Global Label Propagation for Multimodal Fake News Detection

Shuguo Hu, Jun Hu, Huaiwen Zhang

Large Language Models (LLMs) can assist multimodal fake news detection by predicting pseudo labels. However, LLM-generated pseudo labels alone demonstrate poor performance compared to traditional detection methods, making their effective integration non-trivial. In this paper, we propose Global Label Propagation Network with LLM-based Pseudo Labeling (GLPN-LLM) for multimodal fake news detection, which integrates LLM capabilities via label propagation techniques. The global label propagation can utilize LLM-generated pseudo labels, enhancing prediction accuracy by propagating label information among all samples. For label propagation, a mask-based mechanism is designed to prevent label leakage during training by ensuring that training nodes do not propagate their own labels back to themselves. Experimental results on benchmark datasets show that by synergizing LLMs with label propagation, our model achieves superior performance over state-of-the-art baselines.

Are Any-to-Any Models More Consistent Across Modality Transfers Than Specialists?

Jiwan Chung, Janghan Yoon, Junhyeong Park, Sangeyl Lee, Joowon Yang, Sooyeon Park, Youngjae Yu

Any-to-any generative models aim to enable seamless interpretation and generation across multiple modalities within a unified framework, yet their ability to preserve relationships across modalities remains uncertain. Do unified models truly achieve cross-modal coherence, or is this coherence merely perceived? To explore this, we introduce ACON, a dataset of 1,000 images (500 newly contributed) paired with captions, editing instructions, and Q\&A pairs to evaluate cross-modal transfers rigorously. Using three consistency criteria—cyclic consistency, forward equivariance, and conjugated equivariance—our experiments reveal that any-to-any models do not consistently demonstrate greater cross-modal consistency than specialized models in pointwise evaluations such as cyclic consistency. However, equivariance evaluations uncover weak but observable consistency through structured analyses of the intermediate latent space enabled by multiple editing operations. We release our code and data at https://github.com/JiwanChung/ACON.

Con Instruction: Universal Jailbreaking of Multimodal Large Language Models via Non-Textual Modalities

Jiahui Geng, Thy Thy Tran, Preslav Nakov, Iryna Gurevych

Existing attacks against multimodal language models often communicate instruction through text, either as an explicit malicious instruction or a crafted generic prompt, and accompanied by a toxic image. In contrast, here we exploit the capabilities of MLLMs in following non-textual instruction, i.e., an adversarial image or audio, namely Con Instruction. It is a novel gray-box attack method that generates adversarial images or audio to convey specific harmful instructions to MLLMs. We also find that combining our adversarial examples with certain non-empty text inputs amplifies attack success, while appending these after malicious text has limited effects. To evaluate whether an attack is successful, we introduce a new attack response categorization (ARC) that considers the response quality and relevancy concerning the malicious instruction. The results show that Con Instruction effectively bypasses the safety mechanisms in various visual and audio-language models, including LLaVA-v1.5, InternVL, Qwen-VL, and Qwen-Audio, across two standard benchmarks: AdvBench and SafeBench. Specifically, our method achieves the highest attack success rates, reaching 81.3% and 86.6% on LLaVA-v1.5 (13B). We show that larger models are more susceptible toCon Instruction, contrasting observations in their underlying LLMs. On the defense side, we explore various methods against our attacks and find substantial gaps among existing techniques. The code will be made available upon publication.

Grounded, or a Good Guesser? A Per-Question Balanced Dataset to Separate Blind from Grounded Models for Embodied Question Answering

Miles Shelton, Nate Wingerd, Kritim K Rijal, Ayush Garg, Adelina Gutic, Brett Barnes, Catherine Finegan-Dollak

Embodied question answering (EQA) means using *perception of* and *action in* an environment to answer natural language questions about that environment. However, previous work has demonstrated that blind language models (which do not incorporate perception, but predict an answer based solely on the question text) are a strong baseline for existing benchmarks, even compared against state-of-the-art vision and language models. To determine whether a model is grounding its answers in its specific environment, rather than relying on a language model's expectations about the world generally, we propose PQB-EQA, a *per-question balanced* EQA dataset. In this new benchmark, every question appears twice, paired with two different environments that yield two different answers. That is, the answer distribution is balanced for each question, not just across the whole dataset. We show both theoretically and empirically that grounding in the environment is necessary to perform better than chance on PQB-EQA.

Progressive Multimodal Reasoning via Active Retrieval

Guanting Dong, Chenghao Zhang, Mengjie Deng, Yutao Zhu, Zhicheng Dou, Ji-Rong Wen

Multi-step multimodal reasoning tasks pose significant challenges for multimodal large language models (MLLMs), and finding effective ways to enhance their performance in such scenarios remains an unresolved issue. In this paper, we propose AR-MCTS, a universal framework designed to progressively improve the reasoning capabilities of MLLMs through Active Retrieval (AR) and Monte Carlo Tree Search (MCTS). AR-MCTS follows the MCTS algorithm and heuristically integrates an active retrieval mechanism during the expansion stage to automatically acquire high-quality step-wise reasoning annotations. Moreover, we further introduce curriculum training objectives to progressively align with a process reward model, ultimately achieving process-level multimodal reasoning verification. Experimental results across three complex multimodal reasoning benchmarks confirm the effectiveness of AR-MCTS. Further analysis demonstrates that it can optimize sampling diversity and accuracy, yielding reliable multimodal reasoning.

Visual Evidence Prompting Mitigates Hallucinations in Large Vision-Language Models

Wei Li, Zhen Huang, Houqiang Li, Le Lu, Yang Lu, Xinmei Tian, Xu Shen, Jieping Ye

Large Vision-Language Models (LVLMs) have shown impressive progress by integrating visual perception with linguistic understanding to produce contextually grounded outputs. Despite these advancements achieved, LVLMs still suffer from the hallucination problem, e.g., they tend to produce content that does not exist in the input images. Our investigation suggests that such hallucinations often stem from the deficiencies in fine-grained comprehension on the visual aspect, particularly when visual scenes exhibit appearance or semantic similarities (e.g., bicycle vs. motorcycles, baseball bat vs. baseball). In this work, we show such hallucination is naturally mitigated via a novel method called visual evidence prompting, utilizing small visual models to complement the LVLMs. While traditional visual models are not adept at interacting with humans, they excel at perceiving the fine-grained image contents. By symbolizing the professional outputs of domain-expert models as prompts, the LVLM generalists are able to refer to these evidences as visual knowledge to generate more precise answers. Detailed analysis shows that visual evidence enables models to adjust and rectify the attribution and attention on the images, reducing visual confusion by suppressing false activation while enhancing correct ones. Extensive experiments and in-depth analysis demonstrate the effectiveness of our method. We hope our straightforward but insightful work enhances the comprehension of hallucination in LVLMs and offers valuable perspectives on addressing such challenges.

Multimodal Pragmatic Jailbreak on Text-to-image Models

Tong Liu, Zhixin Lai, Jiawen Wang, Gengyuan Zhang, Shuo Chen, Philip Torr, Vera Demberg, Volker Tresp, Jindong Gu

Diffusion models have recently achieved remarkable advancements in terms of image quality and fidelity to textual prompts. Concurrently, the safety of such generative models has become an area of growing concern. This work introduces a novel type of jailbreak, which triggers T2I models to generate the image with visual text, where the image and the text, although considered to be safe in isolation, combine to form unsafe content. To systematically explore this phenomenon, we propose a dataset to evaluate the current diffusion-based text-to-image (T2I) models under such jailbreak. We benchmark nine representative T2I models, including two closed-source commercial models. Experimental results reveal a concerning tendency to produce unsafe content: all tested models suffer from such type of jailbreak, with rates of unsafe generation ranging from around 10% to 70% where DALL·E 3 demonstrates almost the highest unsafety. In real-world scenarios, various filters such as keyword blocklists, customized prompt filters, and NSFW image filters, are commonly employed to mitigate these risks. We evaluate the effectiveness of such filters against our jailbreak and found that, while these filters may be effective for single modality detection, they fail to work against our jailbreak. We also investigate the underlying reason for such jailbreaks, from the perspective of text rendering capability and training data. Our work provides a foundation for further development towards more secure and reliable T2I models.

Mitigating Visual Forgetting via Take-along Visual Conditioning for Multi-modal Long CoT Reasoning

Hai-Long Sun, Zhun Sun, Houwen Peng, Han-Jia Ye

Recent advancements in Large Language Models (LLMs) have demonstrated enhanced reasoning capabilities, evolving from Chain-of-Thought (CoT) prompting to advanced, product-oriented solutions like OpenAI o1. During our re-implementation of this model, we noticed that in multimodal tasks requiring visual input (e.g., geometry problems), Multimodal LLMs (MLLMs) struggle to maintain focus on the visual information, in other words, MLLMs suffer from a gradual decline in attention to visual information as reasoning progresses, causing text-over-relied outputs. To investigate this, we ablate image inputs during long-chain reasoning. Concretely, we truncate the reasoning process midway, then re-complete the reasoning process with the input image removed. We observe only a 2 accuracy drop on MathVista's test-hard subset, revealing the model's textual outputs dominate the following reasoning process. Motivated by this, we propose Take-along Visual Conditioning (TVC), a strategy that shifts image input to critical reasoning stages and compresses redundant visual tokens via dynamic pruning. This methodology helps the model retain attention to the visual components throughout the reasoning. Our approach achieves state-of-the-art performance on average across five mathematical reasoning benchmarks (+3.4% vs previous sota), demonstrating the effectiveness of TVC in enhancing multimodal reasoning systems. The project page is available at .

Can Graph Neural Networks Learn Language with Extremely Weak Text Supervision?

Zihao Li, Lecheng Zheng, Bowen Jin, Dongqi Fu, Baoyu Jing, Yikun Ban, Jingrui He, Jiawei Han

While great success has been achieved in building vision models with Contrastive Language-Image Pre-training (CLIP) over Internet-scale image-text pairs, building transferable Graph Neural Networks (GNNs) with CLIP pipeline is challenging because of the scarcity of labeled data and text supervision, different levels of downstream tasks, and the conceptual gaps between domains. In this work, to address these issues, we propose a multi-modal prompt learning paradigm to effectively adapt pre-trained GNN to downstream tasks and data, given only a few semantically labeled samples, each with extremely weak text supervision. Our new paradigm embeds the graphs directly in the same space as the Large Language Models (LLMs) by learning both graph prompts and text prompts simultaneously. We demonstrate the superior performance of our paradigm in few-shot, multi-task-level, and cross-domain settings. Moreover, we build the first CLIP-style zero-shot classification prototype that can generalize GNNs to unseen classes with extremely weak text supervision.

Redundancy Principles for MLLMs Benchmarks

Zicheng Zhang, Xiangyu Zhao, Xinyu Fang, Chunyi Li, Xiaohong Liu, Xiongkuo Min, Haodong Duan, Kai Chen, Guangtao Zhai

With the rapid iteration of Multi-modality Large Language Models (MLLMs) and the evolving demands of the field, the number of benchmarks produced annually has surged into the hundreds. The rapid growth has inevitably led to significant redundancy among benchmarks. Therefore, it is crucial to take a step back and critically assess the current state of redundancy and propose targeted principles for constructing effective MLLM benchmarks. In this paper, we focus on redundancy from three key perspectives: 1) Redundancy of benchmark capability dimensions, 2) Redundancy in the number of test questions, and 3) Cross-benchmark redundancy within specific domains. Through the comprehensive analysis over hundreds of MLLMs' performance across more than 20 benchmarks, we aim to quantitatively measure the level of redundancy lies in existing MLLM evaluations, provide valuable insights to guide the future development of MLLM benchmarks, and offer strategies to refine and address redundancy issues effectively.

Cultivating Gaming Sense for Yourself: Making VLMs Gaming Experts

wenxuan lu, Jiangyang He, Zhanqiu Zhang, Tianning Zang, Steven Y. Guo

Developing agents capable of fluid gameplay in first/third-person games without API access remains a critical challenge in Artificial General Intelligence (AGI). Recent efforts leverage Vision Language Models (VLMs) as direct controllers, frequently pausing the game to analyze screens and plan action through language reasoning. However, this inefficient paradigm fundamentally restricts agents to basic and non-fluent interactions: relying on isolated VLM reasoning for each action makes it impossible to handle tasks requiring high reactivity (e.g., FPS shooting) or dynamic adaptability (e.g., ACT combat). To handle this, we propose a paradigm shift in gameplay agent design: instead of direct control, VLM serves as a developer, creating specialized execution modules tailored for tasks like shooting and combat. These modules handle real-time game interactions, elevating VLM to a high-level developer. Building upon this paradigm, we introduce GameSense, a gameplay agent framework where VLM develops task-specific game sense modules by observing task execution and leveraging vision tools and neural network training pipelines. These modules encapsulate action-feedback logic, ranging from direct action rules to neural network-based decisions. Experiments demonstrate that our framework is the first to achieve fluent gameplay in diverse genres, including ACT, FPS, and Flappy Bird, setting a new benchmark for game-playing agents.

HiDe-LLaVA: Hierarchical Decoupling for Continual Instruction Tuning of Multimodal Large Language Model

Haiyang Guo, Fanhu Zeng, Ziwei Xiang, Fei Zhu, Da-Han Wang, Xu-Yao Zhang, Cheng-Lin Liu

Instruction tuning is widely used to enhance a pre-trained Multimodal Large Language Model (MLLM) to understand and follow human instructions by training it on a curated set of task-specific dataset. However, it is infeasible to collect all possible instruction datasets simultaneously in real-world scenarios. Thus, enabling MLLM with continual instruction tuning is essential for maintaining their adaptability. However, existing methods often trade off memory efficiency for performance gains, significantly compromising overall efficiency. In this paper, we propose a task-specific expansion and task-general fusion framework based on the variations in Centered Kernel Alignment (CKA) similarity across different model layers when trained on diverse datasets. Furthermore, we analyze the information leakage present in the existing benchmark and propose a new and more challenging benchmark to rationally evaluate the performance of different methods. Comprehensive experiments showcase a significant performance improvement of our method compared to existing state-of-the-art methods. Our code will be public available.

GALLa: Graph Aligned Large Language Models for Improved Source Code Understanding

Ziyin Zhang, Hang Yu, Sage Lee, Peng Di, Jianguo Li, Rui Wang

Programming languages possess rich semantic information - such as data flow - that is represented by graphs and not available from the surface form of source code. Recent code language models have scaled to billions of parameters, but model source code solely as text tokens while ignoring any other structural information. Conversely, models that do encode structural information of code make modifications to the Transformer architecture, limiting their scale and compatibility with pretrained LLMs. In this work, we take the best of both worlds with GALLa - Graph Aligned Large Language Models. GALLa utilizes graph neural networks and cross-modal alignment technologies to inject the structural information of code into LLMs as an auxiliary task during finetuning. This framework is both model-agnostic and task-agnostic, as it can be applied to any code LLM for any code downstream task, and requires the structural graph data only at training time from a corpus unrelated to the finetuning data, while incurring no cost at inference time over the baseline LLM. Experiments on five code tasks with six different baseline LLMs ranging in size from 350M to 14B validate the effectiveness of GALLa, demonstrating consistent improvement over the baseline, even for powerful models such as LLaMA3 and Qwen2.5-Coder.

Investigating and Enhancing the Robustness of Large Multimodal Models Against Temporal Inconsistency

Jiafeng Liang, Shixin Jiang, Xuan Dong, Ning Wang, Zheng Chu, Hui Su, Jinlan Fu, Ming Liu, See-Kiong Ng, Bing Qin

Large Multimodal Models (LMMs) have recently demonstrated impressive performance on general video comprehension benchmarks. Nevertheless, for broader applications, the robustness of their temporal analysis capability needs to be thoroughly investigated yet predominantly ignored. Motivated by this, we propose a novel temporal robustness benchmark (TemRobBench), which introduces temporal inconsistency perturbations separately at the visual and textual modalities to assess the robustness of models. We evaluate 16 mainstream LMMs and find that they exhibit over-reliance on prior knowledge and textual context in adversarial environments, while ignoring the actual temporal dynamics in the video. To mitigate this issue, we design panoramic direct preference optimization (PanoDPO), which encourages LMMs to incorporate both visual and linguistic feature preferences simultaneously. Experimental results show that PanoDPO can effectively enhance the model's robustness and reliability in temporal analysis.

Contextual Experience Replay for Continual Learning of Language Agents

Yitao Liu, Chenglei Si, Karthik R Narasimhan, Shunyu Yao

Large language model (LLM) agents have been applied to sequential decision-making tasks such as web navigation, but without any environment-specific experiences, they often fail in these complex tasks. Moreover, current LLM agents are not designed to continually learn from past experiences during inference time, which could be crucial for them to gain these environment-specific experiences. To address this, we propose Contextual Experience Replay (CER), a training-free framework to enable efficient self-improvement for language agents in their context window. Specifically, CER accumulates and synthesizes past experiences into a dynamic memory buffer. These experiences encompass environment dynamics and common decision-making patterns, allowing the agents to retrieve and augment themselves with relevant knowledge in new tasks, enhancing their adaptability in complex environments. We evaluate CER on the challenging WebArena and VisualWebArena benchmarks. On VisualWebArena, CER surpasses the tree search method with much fewer token costs and achieves the state-of-the-art performance of 31.9%. On WebArena, CER also gets a competitive average success rate of 36.7%, relatively improving the success rate of the GPT-4o agent baseline by 51.0%. We also conduct a comprehensive analysis on it to prove its efficiency, validity and understand it better.

Emma-X: An Embodied Multimodal Action Model with Grounded Chain of Thought and Look-ahead Spatial Reasoning

Qi Sun, Pengfei Hong, Tej Deep Pala, Vernon Toh, U-Xuan Tan, Deepanway Ghosal, Soujanya Poria

Traditional reinforcement learning-based robotic control methods are often task-specific and fail to generalize across diverse environments or unseen objects and instructions. Visual Language Models (VLMs) demonstrate strong scene understanding and planning capabilities but lack the ability to generate actionable policies tailored to specific robotic embodiments. To address this, Visual-Language-Action (VLA) models have emerged, yet they face challenges in long-horizon spatial reasoning and grounded task planning. In this work, we propose the Embodied Multimodal Action Model with Grounded Chain of Thought and Look-ahead Spatial Reasoning, EMMA-X. EMMA-X leverages our constructed hierarchical embodiment dataset based on BridgeV2, containing 60,000 robot manipulation trajectories auto-annotated with grounded task reasoning and spatial guidance. Additionally, we introduce a trajectory segmentation strategy based on gripper states and motion trajectories, which can help mitigate hallucination in grounding subtask reasoning generation. Experimental results demonstrate that EMMA-X achieves superior performance over competitive baselines, particularly in real-world robotic tasks requiring spatial reasoning.

Browsing Like Human: A Multimodal Web Agent with Experiential Fast-and-Slow Thinking

Haohao Luo, Jiayi Kuang, Wei Liu, Ying Shen, Jian Luan, Yang Deng

Automating web navigation which aims to build a web agent that follows user instructions to complete tasks like booking flights by interacting with websites, has received increasing attention due to its practical value. Although existing web agents are mostly equipped with visual perception, planning, and memory abilities, their reasoning process are still deviate from human cognition. In this work, we study the human thought pattern to empower agent with more human-like abilities in web navigation. To tackle this problem, we propose a novel multimodal web agent framework called WebExperT, which is designed to emulate the human planning process of "thinking fast and slow" to effectively decompose complex user instructions. Furthermore, WebExperT leverages experiential learning by reflecting from failure for continuously refining planning and decision-making outcomes. Experimental results on the Mind2Web benchmark demonstrate the superiority of WebExperT in both supervised and unsupervised settings.

Can MLLMs Understand the Deep Implication Behind Chinese Images?

Chenhao Zhang, Xi Feng, Yuelin Bai, Xeron Du, Jinchang Hou, Kaixin Deng, Guangzeng Han, Qinrui Li, Bingli Wang, Jiaheng Liu, Xingwei Qu, Yifei Zhang, Qixuan Zhao, Yiming Liang, Ziqiang Liu, Feiteng Fang, Min Yang, Wenhao Huang, Chenghua Lin, Ge Zhang, Shiwen Ni

As the capabilities of Multimodal Large Language Models (MLLMs) improve, the need for higher-order evaluation of them is increasing. However, there is a lack of work evaluating MLLM for higher-order perception and understanding of Chinese visual content. To address this, we introduce the CII-Bench, which aims to assess MLLMs' such capabilities for Chinese images. To ensure the authenticity of the Chinese context, images in CII-Bench are sourced from the Chinese Internet and manually reviewed, with corresponding answers also manually crafted. Additionally, CII-Bench incorporates images that represent Chinese traditional culture, such as famous Chinese traditional paintings, which can deeply reflect the model's understanding of Chinese traditional culture. Through experiments on multiple MLLMs using CII-Bench, significant findings emerged. There is a large gap between MLLMs and humans in performance. The highest MLLM accuracy is 64.4%, while the human average is 78.2% and the peak is 81.0%. MLLMs perform poorly on traditional culture images, indicating limitations in understanding high-level semantics and lacking a deep knowledge base of Chinese traditional culture. Moreover, most models have higher accuracy when image emotion hints are added to the prompts. We believe CII-Bench will help MLLMs better understand Chinese semantics and specific images, and move forward the development of expert artificial general intelligence (AGI). Our project is publicly available at https://cii-bench.github.io.

EAGLE: Expert-Guided Self-Enhancement for Preference Alignment in Pathology Large Vision-Language Model

Meidan Ding, Jipeng Zhang, Wenxuan Wang, Haiqin Zhong, Xiaoqin Wang, XINHENG LYU, Wenting Chen, Linlin Shen

Recent advancements in Large Vision Language Models (LVLMs) show promise for pathological diagnosis, yet their application in clinical settings faces critical challenges of multimodal hallucination and biased responses. While preference alignment methods have proven effective in general domains, acquiring high-quality preference data for pathology remains challenging due to limited expert resources and domain complexity. In this paper, we propose EAGLE (Expert-guided self-enhancement for preference Alignment in patholoGy Large vision-languagE model), a novel framework that systematically integrates medical expertise into preference alignment. EAGLE consists of three key stages: initialization through supervised fine-tuning, self-preference creation leveraging expert prompting and medical entity recognition, and iterative preference following-tuning. The self-preference creation stage uniquely combines expert-verified chosen sampling with expert-guided rejected sampling to generate high-quality preference data, while the iterative tuning process continuously refines both data quality and model performance. Extensive experiments demonstrate that EAGLE significantly outperforms existing pathological LVLMs, effectively reducing hallucination and bias while maintaining pathological accuracy. The source code is available at https://github.com/meidandz/EAGLE.

HiddenDetect: Detecting Jailbreak Attacks against Multimodal Large Language Models via Monitoring Hidden States

Yilei Jiang, Xinyan Gao, Tianshuo Peng, Yingshui Tan, Xiaoyong Zhu, Bo Zheng, Xiangyu Yue

The integration of additional modalities increases the susceptibility of large vision-language models (LVLMs) to safety risks, such as jailbreak attacks, compared to their language-only counterparts. While existing research primarily focuses on post-hoc alignment techniques, the underlying safety mechanisms within LVLMs remain largely unexplored. In this work , we investigate whether LVLMs inherently encode safety-relevant signals within their internal activations during inference. Our findings reveal that LVLMs exhibit distinct activation patterns when processing unsafe prompts, which can be leveraged to detect and mitigate adversarial inputs without requiring extensive fine-tuning. Building on this insight, we introduce HiddenDetect, a novel tuning-free framework that harnesses internal model activations to enhance safety. Experimental results show that {HiddenDetect} surpasses state-of-the-art methods in detecting jailbreak attacks against LVLMs. By utilizing intrinsic safety-aware patterns, our method provides an efficient and scalable solution for strengthening LVLM robustness against multimodal threats. Our code and data will be released publicly.

World Modeling Makes a Better Planner: Dual Preference Optimization for Embodied Task Planning

Siyin Wang, Zhaoye Fei, Qinyuan Cheng, Shiduo Zhang, Panpan Cai, Jinlan Fu, Xipeng Qiu

Recent advances in large vision-language models (LVLMs) have shown promise for embodied task planning, yet they struggle with fundamental challenges like dependency constraints and efficiency. Existing approaches either solely optimize action selection or directly leverage pre-trained models as world models during inference, overlooking the benefits of learning to model the world as a way to enhance planning capabilities. We propose Dual Preference Optimization (D^2PO), a new learning framework that jointly optimizes state prediction and action selection through preference learning, enabling LVLMs to understand environment dynamics for better planning. To automatically collect trajectories and stepwise preference data without human annotation, we introduce a tree search mechanism for extensive exploration via trial-and-error. Extensive experiments on VoTa-Bench demonstrate that our D^2PO-based method significantly outperforms existing methods and GPT-4o when applied to Qwen2-VL (7B), LLaVA-1.6 (7B), and LLaMA-3.2 (11B), achieving superior task success rates with more efficient execution paths.

VisuoThink: Empowering LVLM Reasoning with Multimodal Tree Search

Yikun Wang, Siyin Wang, Qinyuan Cheng, Zhaoye Fei, Liang Ding, Qipeng Guo, Dacheng Tao, Xipeng Qiu

Recent advancements in Large Vision-Language Models have showcased remarkable capabilities. However, they often falter when confronted with complex reasoning tasks that humans typically address through visual aids and deliberate, step-by-step thinking. While existing methods have explored text-based slow thinking or rudimentary visual assistance, they fall short of capturing the intricate, interleaved nature of human visual-verbal reasoning processes. To overcome these limitations and inspired by the mechanisms of slow thinking in human cognition, we introduce VisuoThink, a novel framework that seamlessly integrates visuospatial and linguistic domains. VisuoThink facilitates multimodal slow thinking by enabling progressive visual-textual reasoning and incorporates test-time scaling through look-ahead tree search. Extensive experiments demonstrate that VisuoThink significantly enhances reasoning capabilities via inference-time scaling, even without fine-tuning, achieving state-of-the-art performance in tasks involving geometry and spatial reasoning.

MMRC: A Large-Scale Benchmark for Understanding Multimodal Large Language Model in Real-World Conversation

Haochen Xue, Feilong Tang, Ming Hu, Yexin Liu, Qidong Huang, Yulong Li, Chengzhi Liu, Zhongxing Xu, Chong Zhang, Chun-Mei Feng, Yutong Xie, Imran Razzak, Zongyuan Ge, Jionglong Su, Junjun He, Yu Qiao

Recent multimodal large language models (MLLMs) have demonstrated significant potential in open-ended conversation, generating more accurate and personalized responses. However, their abilities to memorize, recall, and reason in sustained interactions within real-world scenarios remain underexplored. This paper introduces MMRC, a Multi-Modal Real-world Conversation benchmark for evaluating six core open-ended abilities of MLLMs: information extraction, multi-turn reasoning, information update, image management, memory recall, and answer refusal. With data collected from real-world scenarios, MMRC comprises 5,120 conversations and 28,720 corresponding manually labeled questions, posing a significant challenge to existing MLLMs. Evaluations on 20 MLLMs in MMRC indicate an accuracy drop during open-ended interactions. We identify four common failure patterns: long-term memory degradation, inadequacies in updating factual knowledge, accumulated assumption of error propagation, and reluctance to "say no." To mitigate these issues, we propose a simple yet effective NOTE-TAKING strategy, which can record key information from the conversation and remind the model during its responses, enhancing conversational capabilities. Experiments across six MLLMs demonstrate significant performance improvements.

FCMR: Robust Evaluation of Financial Cross-Modal Multi-Hop Reasoning

Seunghee Kim, Changhyeon Kim, Taeuk Kim

Real-world decision-making often requires integrating and reasoning over information from multiple modalities. While recent multimodal large language models (MLLMs) have shown promise in such tasks, their ability to perform multi-hop reasoning across diverse sources remains insufficiently evaluated. Existing benchmarks, such as MMQA, face challenges due to (1) data contamination and (2) a lack of complex queries that necessitate operations across more than two modalities, hindering accurate performance assessment. To address this, we present Financial Cross-Modal Multi-Hop Reasoning (FCMR), a benchmark created to analyze the reasoning capabilities of MLLMs by urging them to combine information from textual reports, tables, and charts within the financial domain. FCMR is categorized into three difficulty levels---Easy, Medium, and Hard---facilitating a step-by-step evaluation. In particular, problems at the Hard level require precise cross-modal three-hop reasoning and are designed to prevent the disregard of any modality. Experiments on this new benchmark reveal that even state-of-the-art MLLMs struggle, with the best-performing model (Claude 3.5 Sonnet) achieving only 30.4% accuracy on the most challenging tier. We also conduct analysis to provide insights into the inner workings of the models, including the discovery of a critical bottleneck in the information retrieval phase.

Sparse-to-Dense: A Free Lunch for Lossless Acceleration of Video Understanding in LLMs

Xuan Zhang, Cunxiao Du, Sicheng Yu, Jiawei Wu, Fengzhuo Zhang, Wei Gao, Qian Liu

Due to the auto-regressive nature of current video large language models (Video-LLMs), the inference latency increases as the input sequence length grows, posing challenges for the efficient processing of video sequences that are usually very long. We observe that during decoding, the attention scores of most tokens in Video-LLMs tend to be sparse and concentrated, with only certain tokens requiring comprehensive full attention. Based on this insight, we introduce Sparse-to-Dense (StD), a novel decoding strategy that integrates two distinct modules: one leveraging sparse top-K attention and the other employing dense full attention. These modules collaborate to accelerate Video-LLMs without loss. The fast (sparse) model speculatively decodes multiple tokens, while the slow (dense) model verifies them in parallel. StD is a tuning-free, plug-and-play solution that achieves up to a 1.94 walltime speedup in video processing. It maintains model performance while enabling a seamless transition from a standard Video-LLM to a sparse Video-LLM with minimal code modifications.

OMGM: Orchestrate Multiple Granularities and Modalities for Efficient Multimodal Retrieval

Wei Yang, Jingjing Fu, Rui Wang, Jinyu Wang, Lei Song, Jiang Bian

Vision-language retrieval-augmented generation (RAG) has become an effective approach for tackling Knowledge-Based Visual Question Answering (KB-VQA), which requires external knowledge beyond the visual content presented in images. The effectiveness of Vision-language RAG systems hinges on multimodal retrieval, which is inherently challenging due to the diverse modalities and knowledge granularities in both queries and knowledge bases. Existing methods have not fully tapped into the potential interplay between these elements. We propose a multimodal RAG system featuring a coarse-to-fine, multi-step retrieval that harmonizes multiple granularities and modalities to enhance efficacy. Our system begins with a broad initial search aligning knowledge granularity for cross-modal retrieval, followed by a multimodal fusion reranking to capture the nuanced multimodal information for top entity selection. A text reranker then filters out the most relevant fine-grained section for augmented generation. Extensive experiments on the InfoSeek and Encyclopedic-VQA benchmarks show our method achieves state-of-the-art retrieval performance and highly competitive answering results, underscoring its effectiveness in advancing KB-VQA systems. Our code can be found at https://github.com/ChaoLinAViy/OMGM.

Can LLMs Deceive CLIP? Benchmarking Adversarial Compositionality of Pre-trained Multimodal Representation via Text Updates

Jaewoo Ahn, Heeseung Yun, Dayoon Ko, Gunhee Kim

While pre-trained multimodal representations (e.g., CLIP) have shown impressive capabilities, they exhibit significant compositional vulnerabilities leading to counterintuitive judgments. We introduce Multimodal Adversarial Compositionality (MAC), a benchmark that leverages large language models (LLMs) to generate deceptive text samples to exploit these vulnerabilities across different modalities and evaluates them through both sample-wise attack success rate and group-wise entropy-based diversity. To improve zero-shot methods, we propose a self-training approach that leverages rejection-sampling fine-tuning with diversity-promoting filtering, which enhances both attack success rate and sample diversity. Using smaller language models like Llama-3.1-8B, our approach demonstrates superior performance in revealing compositional vulnerabilities across various multimodal representations, including images, videos, and audios.

LPOI: Listwise Preference Optimization for Vision Language Models

Fatemeh Pesaran zadeh, Yoojin Oh, Gunhee Kim

Aligning large VLMs with human preferences is a challenging task, as methods like RLHF and DPO often overfit to textual information or exacerbate hallucinations. Although augmenting negative image samples partially addresses these pitfalls, no prior work has employed listwise preference optimization for VLMs, due to the complexity and cost of constructing listwise image samples. In this work, we propose LPOI, the first object-aware listwise preference optimization developed for reducing hallucinations in VLMs. LPOI identifies and masks a critical object in the image, and then interpolates the masked region between the positive and negative images to form a sequence of incrementally more complete images. The model is trained to rank these images in ascending order of object visibility, effectively reducing hallucinations while retaining visual fidelity. LPOI requires no extra annotations beyond standard pairwise preference data, as it automatically constructs the ranked lists through object masking and interpolation. Comprehensive experiments on MMHalBench, AMBER, and Object HalBench confirm that LPOI outperforms existing preference optimization methods in reducing hallucinations and enhancing VLM performance.

Walk in Others' Shoes with a Single Glance: Human-Centric Visual Grounding with Top-View Perspective Transformation

Yuqi Bu, Xin Wu, Zirui Zhao, Yi Cai, David Hsu, Qiong Liu

Visual perspective-taking, an ability to envision others' perspectives from a single self-perspective, is vital in human-robot interactions. Thus, we introduce a human-centric visual grounding task and a dataset to evaluate this ability. Recent advances in vision-language models (VLMs) have shown potential for inferring others' perspectives, yet are insensitive to information differences induced by slight perspective changes. To address this problem, we propose a top-view enhanced perspective transformation (TEP) method, which decomposes the transition from robot to human perspectives through an abstract top-view representation. It unifies perspectives and facilitates the capture of information differences from diverse perspectives. Experimental results show that TEP improves performance by up to 18%, exhibits perspective-taking abilities across various perspectives, and generalizes effectively to robotic and dynamic scenarios.

Transferring Textual Preferences to Vision-Language Understanding through Model Merging

Chen-An Li, Tzu-Han Lin, Yun-Nung Chen, Hung-yi Lee

Large vision-language models (LVLMs) perform outstandingly across various multimodal tasks. However, their ability to evaluate generated content remains limited, and training vision-language reward models (VLRMs) with preference data is computationally expensive. This paper explores a training-free alternative by merging text-based reward models (RMs) with LVLMs to create VLRMs. Our approach shows that integrating these models leads to improved performance over LVLMs' scoring and text-based RMs, offering an efficient method for incorporating textual preferences into LVLMs.

FinMME: A Financial Multi-Modal Evaluation Dataset

Junyu Luo, Zhizhuo KOU, Liming Yang, Xiao Luo, Jinsheng Huang, Zhiping Xiao, Jingshu Peng, Chengzhong LIU, Jiaming Ji, Xuanzhe Liu, Sirui Han, Ming Zhang, Yike Guo

Multimodal Large Language Models (MLLMs) have experienced rapid development in recent years. However, in the financial domain, there is a notable lack of effective and specialized multimodal evaluation datasets. To advance the development of MLLMs in the finance domain, we introduce FinMME, encompassing more than 11,000 high-quality financial research samples across 18 financial domains and 6 asset classes, featuring 10 major chart types and 21 subtypes. We ensure data quality through 20 annotators and carefully designed validation mechanisms. Additionally, we develop FinScore, an evaluation system incorporating hallucination penalties and multi-dimensional capability assessment to provide an unbiased evaluation. Extensive experimental results demonstrate that even state-of-the-art models like GPT-4o exhibit unsatisfactory performance on FinMME, highlighting its challenging nature. The benchmark exhibits high robustness with prediction variations under different prompts remaining below 1\%, demonstrating superior reliability compared to existing datasets. Our dataset and evaluation protocol are available at https://huggingface.co/datasets/luojunyu/FinMME and https://github.com/luo-junyu/FinMME.

METAL: A Multi-Agent Framework for Chart Generation with Test-Time Scaling

Bingxuan Li, Yiwei Wang, Jiuxiang Gu, Kai-Wei Chang, Nanyun Peng

Chart generation aims to generate code to produce charts satisfying the desired visual properties, e.g., texts, layout, color, and type. It has great potential to empower the automatic professional report generation in financial analysis, research presentation, education, and healthcare. In this work, we build a vision-language model (VLM) based multi-agent framework for effective automatic chart generation. Generating high-quality charts requires both strong visual design skills and precise coding capabilities that embed the desired visual properties into code. Such a complex multi-modal reasoning process is difficult for direct prompting of VLMs. To resolve these challenges, we propose METAL, a multi-agent framework that decomposes the task of chart generation into the iterative collaboration among specialized agents. METAL achieves a 5.2\% improvement in the F1 score over the current best result in the chart generation task. Additionally, METAL improves chart generation performance by 11.33% over Direct Prompting with LLaMA-3.2-11B. Furthermore, the METAL framework exhibits the phenomenon of test-time scaling: its performance increases monotonically as the logarithm of computational budget grows from 512 to 8192 tokens.

CLIPErase: Efficient Unlearning of Visual-Textual Associations in CLIP

Tianyu Yang, Lisen Dai, Xiangqi Wang, Minhao Cheng, Yapeng Tian, Xiangliang Zhang

Machine unlearning (MU) has gained significant attention as a means to remove the influence of specific data from a trained model without requiring full retraining. While progress has been made in unimodal domains like text and image classification, unlearning in multimodal models remains relatively under-explored. In this work, we address the unique challenges of unlearning in CLIP, a prominent multimodal model that aligns visual and textual representations. We introduce CLIPErase, a novel approach that disentangles and selectively forgets both visual and textual associations, ensuring that unlearning does not compromise model performance. CLIPErase consists of three key modules: a Forgetting Module that disrupts the associations in the forget set, a Retention Module that preserves performance on the retain set, and a Consistency Module that maintains consistency with the original model. Extensive experiments on CIFAR-100, Flickr30K, and Conceptual 12M across five CLIP downstream tasks, as well as an evaluation on diffusion models, demonstrate that CLIPErase effectively removes designated associations from multimodal samples in downstream tasks, while preserving the model's performance on the retain set after unlearning.

ISR: Self-Refining Referring Expressions for Entity Grounding

Zhuocheng Yu, Bingchan Zhao, Yifan Song, Sujian Li, ZHONGHUI HE

Entity grounding, a crucial task in constructing multimodal knowledge graphs, aims to align entities from knowledge graphs with their corresponding images. Unlike conventional visual grounding tasks that use referring expressions (REs) as inputs, entity grounding relies solely on entity names and types, presenting a significant challenge. To address this, we introduce a novel **I**terative **S**elf-**R**efinement (**ISR**) scheme to enhance the multimodal large language model's capability to generate high quality REs for the given entities as explicit contextual clues. This training scheme, inspired by human learning dynamics and human annotation processes, enables the MLLM to iteratively generate and refine REs by learning from successes and failures, guided by outcome rewards from a visual grounding model. This iterative cycle of self-refinement avoids overfitting to fixed annotations and fosters continued improvement in referring expression generation. Extensive experiments demonstrate that our methods surpasses other methods in entity grounding, highlighting its effectiveness, robustness and potential for broader applications.

Multi-Modality Expansion and Retention for LLMs through Parameter Merging and Decoupling

Junlin Li, Guodong DU, Jing Li, Sim Kuan Goh, Wenya Wang, Yequan Wang, Fangming Liu, Ho-Kin Tang, Saleh Alharbi, Daojing He, Min Zhang

Fine-tuning Large Language Models (LLMs) with multimodal encoders on modality-specific data expands the modalities that LLMs can handle, leading to the formation of Multimodal LLMs (MLLMs). However, this paradigm heavily relies on resource-intensive and inflexible fine-tuning from scratch with new multimodal data. In this paper, we propose MMER (Multi-modality Expansion and Retention), a training-free approach that integrates existing MLLMs for effective multimodal expansion while retaining their original performance. Specifically, MMER reuses MLLMs' multimodal encoders while merging their LLM parameters. By comparing original and merged LLM parameters, MMER generates binary masks to approximately separate LLM parameters for each modality. These decoupled parameters can independently process modality-specific inputs, reducing parameter conflicts and preserving original MLLMs' fidelity. MMER can also mitigate catastrophic forgetting by applying a similar process to MLLMs fine-tuned on new tasks. Extensive experiments show significant improvements over baselines, proving that MMER effectively expands LLMs' multimodal capabilities while retaining 99% of the original performance, and also markedly mitigates catastrophic forgetting.

CityNavAgent: Aerial Vision-and-Language Navigation with Hierarchical Semantic Planning and Global Memory

Weichen Zhang, Chen Gao, Shiquan Yu, Ruiying Peng, Baining Zhao, Qian Zhang, Jinqiang Cui, Xinlei Chen, Yong Li

Aerial vision-and-language navigation (VLN) — requiring drones to interpret natural language instructions and navigate complex urban environments — emerges as a critical embodied AI challenge that bridges human-robot interaction, 3D spatial reasoning, and real-world deployment. Although existing ground VLN agents achieved notable results in indoor and outdoor settings, they struggle in aerial VLN due to the absence of predefined navigation graphs and the exponentially expanding action space in long-horizon exploration. In this work, we propose \textbf{CityNavAgent}, a large language model (LLM)-empowered agent that significantly reduces the navigation complexity for urban aerial VLN. Specifically, we design a hierarchical semantic planning module (HSPM) that decomposes the long-horizon task into sub-goals with different semantic levels. The agent reaches the target progressively by achieving sub-goals with different capacities of the LLM. Additionally, a global memory module storing historical trajectories into a topological graph is developed to simplify navigation for visited targets. Extensive benchmark experiments show that our method achieves state-of-the-art performance with significant improvement. Further experiments demonstrate the effectiveness of different modules of CityNavAgent for aerial VLN in continuous city environments.

Maximal Matching Matters: Preventing Representation Collapse for Robust Cross-Modal Retrieval

Hani Alomari, Anushka Sivakumar, Andrew Zhang, Chris Thomas

Cross-modal image-text retrieval is challenging because of the diverse possible associations between content from different modalities. Traditional methods learn a single-vector embedding to represent semantics of each sample, but struggle to capture nuanced and diverse relationships that can exist across modalities. Set-based approaches, which represent each sample with multiple embeddings, offer a promising alternative, as they can capture richer and more diverse relationships. In this paper, we show that, despite their promise, these set-based representations continue to face issues including sparse supervision and set collapse, which limits their effectiveness. To address these challenges, we propose Maximal Pair Assignment Similarity to optimize one-to-one matching between embedding sets which preserve semantic diversity within the set. We also introduce two loss functions to further enhance the representations: Global Discriminative Loss to enhance distinction among embeddings, and Intra-Set Divergence Loss to prevent collapse within each set. Our method achieves state-of-the-art performance on MS-COCO and Flickr30k without relying on external data.

Scalable Vision Language Model Training via High Quality Data Curation

Hongyuan Dong, Zijian Kang, Weijie Yin, LiangXiao, ChaoFeng, Ran Jiao
In this paper, we introduce \textit{\textbf{SAIL}-VL} (\textit{\textbf{S}c\textbf{A}lable Vision Language Model Tra\textbf{I}ning via High Qua\textbf{L}ity Data Curation}), an open-source vision language model (VLM) series achieving state-of-the-art (SOTA) performance in 2B and 8B parameters. The following three key improvements contribute to SAIL-VL's leading performance: (1) Scalable high-quality visual understanding data construction: We implement a data construction pipeline to enable hundred-million-scale high-quality recaption data annotation. The resulted dataset SAIL-Caption is validated to be of the highest data quality compared with opensource datasets. (2) Scalable Pretraining with High-Quality Visual Understanding Data: We scale SAIL-VL's pretraining budget up to 655B tokens and show that even a 2B VLM benefits from scaled up training data sizes, exhibiting logarithmic data size scaling laws in benchmark performance. (3) Scalable SFT via data quantity and complexity scaling: We curate a high-quality SFT dataset collection with leading data quantity scaling effectiveness and demonstrate that training with progressively higher-complexity data surpasses baseline one-stage training by a large margin. SAIL-VL series models achieve the highest average score in 18 widely used VLM benchmarks in our evaluation, with the 2B model takes the top position over VLMs of comparable sizes on OpenCompass 2024 (https://rank.opencompass.org.cn/leaderboard-multimodal), demonstrating robust visual comprehension abilities. SAIL-VL series models are released at HuggingFace (https://huggingface.co/BytedanceDouyinContent).

Design Choices for Extending the Context Length of Visual Language Models

Mukai Li, Lei Li, Shansan Gong, Qi Liu
Visual Language Models (VLMs) demonstrate impressive capabilities in processing multimodal inputs, yet applications such as visual agents, which require handling multiple images and high-resolution videos, demand enhanced long-range modeling. Moreover, existing open-source VLMs lack systematic exploration into extending their context length, and commercial models often provide limited details. To tackle this, we aim to establish an effective solution that enhances long context performance of VLMs while preserving their capacities in short context scenarios. Towards this goal, we make the best design choice through extensive experiment settings from data curation to context window extending and utilizing: (1) we analyze data sources and length distributions to construct ETVLM - a data recipe to balance the performance across scenarios; (2) we examine existing position extending methods, identify their limitations and propose M-RoPE++ as an enhanced approach; we also choose to solely instruction-tune the backbone with mixed-source data; (3) we discuss how to better utilize extended context windows and propose hybrid-resolution training. Built on the Qwen-VL series model, we propose Giraffe, which is effectively extended to 128K lengths. Evaluated on extensive long context VLM benchmarks such as VideoMME and Viusal Haystacks, our Giraffe achieves state-of-the-art performance among similarly sized open-source long VLMs and is competitive with commercial model GPT-4V. We will open-source the code, data, and models.

Session 10: Multimodality and Language Grounding to Vision, Robotics and Beyond (Tuesday 29 July, 16:00-17:30, Hall 5X)

DEEP: an automatic bidirectional translator leveraging an ASR for translation of Italian sign language

Nicolas Tagliabue, Elisa Colletti, Francesco Roberto Dani, Roberto Tedesco, Sonia Cenceschi, Alessandro Trivilini
DEEP is a bidirectional translation system for the Italian Sign Language, tailored to two specific, common use cases: pharmacy and the registry office of the municipality, for which a custom corpus has been collected. Two independent pipelines permit to create a chat-like interaction style, where the deaf subject just signs in front of a camera, and see a virtual LIS interpreter, while the hearing subject read and write messages into a chat UI. The LIS-to-Italian pipeline leverages, in a novel way, a customized Whisper model (a well-known speech recognition system), by means of "pseudo-spectrograms". The Italian-to-LIS pipeline leverages a virtual avatar created with Viggle.ai. DEEP has been evaluated with LIS signers, obtaining very promising results.

FORG3D: Flexible Object Rendering for Generating Vision-Language Spatial Reasoning Data from 3D Scenes

Oscar Pang, Freda Shi

We introduce FORG3D, a 3D rendering toolkit developed with Blender and Python, which synthesizes vision-language data for two primary purposes: (1) supporting human cognitive experiments that require fine-grained control over material and (2) analyzing and improving the visual reasoning capabilities of large vision-language models. The toolkit provides flexible and precise control over object placement, orientation, inter-object distances, and camera configurations while automatically generating detailed spatial metadata. Additionally, it includes a built-in feature for integrating AI-generated backgrounds, enhancing the realism of synthetic scenes. FORG3D is publicly available at https://github.com/compling-wat/FORG3D, and a video demonstration is available at https://www.youtube.com/watch?v=QvIqib_PU8A.

LECTURE4ALL: A Lightweight Approach to Precise Timestamp Detection in Online Lecture Videos

Viktoria Wrobel, Simon Kazemi, Frank Hammerschmidt, Torben Hannemann, Gregor Stange, Seid Muhie Yimam, Robert Geislinger

This paper presents LECTURE4ALL, a web application developed to improve the search experience of educational video platforms. Lecture2Go provides a vast collection of recorded lectures, but locating specific content within videos can be time-consuming. LEC- TURE4ALL addresses this issue by leveraging a vector database and a streamlined user inter- face to enable direct retrieval of precise video timestamps. By enhancing search accuracy and efficiency, LECTURE4ALL significantly im- proves the accessibility and usability of educa- tional video platforms.

Session 10: Resources and Evaluation (Tuesday 29 July, 16:00-17:30, Hall 5X)

CodeArena: A Collective Evaluation Platform for LLM Code Generation

Mingzhe Du, Anh Tuan Luu, Bin Ji, Xiaobao Wu, Yuhao QING, Dong HUANG, Terry Yue Zhuo, Qian Liu, See-Kiong Ng

Large Language Models (LLMs) have reshaped code generation by synergizing their exceptional comprehension of natural language and programming syntax, thereby substantially boosting developer productivity. These advancements have prompted numerous efforts to quantitatively evaluate their coding capabilities. However, persistent challenges, such as benchmark leakage, data dissipation, and limited system accessibility, continue to impede a timely and accurate assessment. To address these limitations, we introduce CodeArena, an online evaluation framework tailored for LLM code generation. Its key innovation is a collective evaluation mechanism, which dynamically recalibrates individual model scores based on the holistic performance of all participating models, mitigating score biases caused by widespread benchmark leakage. In addition, CodeArena ensures open access to all submitted solutions and test cases and provides automation-friendly APIs to streamline the code evaluation workflow. Our main contributions are: (1) a collective evaluation system for unbiased assessment, (2) a public repository of solutions and test cases, and (3) automation-ready APIs for seamless integration.

gec-metrics: A Unified Library for Grammatical Error Correction Evaluation

Takumi Goto, Yusuke Sakai, Taro Watanabe

We introduce gec-metrics, a library for using and developing grammatical error correction (GEC) evaluation metrics through a unified interface. Our library enables fair system comparisons by ensuring that everyone conducts evaluations using a consistent implementation. Moreover, it is designed with a strong focus on API usage, making it highly extensible. It also includes meta-evaluation functionalities and provides analysis and visualization scripts, contributing to developing GEC evaluation metrics. Our code is released under the MIT license^1 and is also distributed as an installable package^2. The video is available at YouTube^3. ^1GitHub: https://github.com/gotutiyan/gec-metrics ^2PyPi: https://pypi.org/project/gec-metrics/ ^3Video: https://youtu.be/cor6dkN6EfI

FlagEval-Arena: A Side-by-Side Comparative Evaluation Platform for Large Language Models and Text-Driven AIGC

Jing-Shu Zheng, richeng xuan, Bowen Qin, Zheqi He, Tongshuai.ren, Xuejing Li, Jin-Ge Yao, Xi Yang

We introduce FlagEval-Arena, an evaluation platform for side-by-side comparisons of large language models and text-driven AIGC systems.Compared with the well-known LM Arena (LMSYS Chatbot Arena), we reimplement our own framework with the flexibility to introduce new mechanisms or features. Our platform enables side-by-side evaluation not only for language models or vision-language models, but also text-to-image or text-to-video synthesis. We specifically target at Chinese audience with a more focus on the Chinese language, more models developed by Chinese institutes, and more general usage beyond the technical community. As a result, we currently observe very interesting differences from usual results presented by LM Arena. Our platform is available via this URL: .

Session 10: Special Theme: Generalization of NLP Models (Tuesday 29 July, 16:00-17:30, Hall 4/5)

Learning Sparsity for Effective and Efficient Music Performance Question Answering

Xingjian Diao, Tianzhen Yang, Chunhui Zhang, Weiyi Wu, Ming Cheng, Jiang Gui

Music performances, characterized by dense and continuous audio as well as seamless audio-visual integration, present unique challenges for multimodal scene understanding and reasoning. Recent Music Performance Audio-Visual Question Answering (Music AVQA) datasets have been proposed to reflect these challenges, highlighting the continued need for more effective integration of audio-visual representations in complex question answering. However, existing Music AVQA methods often rely on dense and unoptimized representations, leading to inefficiencies in the isolation of key information, the reduction of redundancy, and the prioritization of critical samples. To address these challenges, we introduce Sparsify, a sparse learning framework specifically designed for Music AVQA. It integrates three sparsification strategies into an end-to-end pipeline and achieves state-of-the-art performance on the Music AVQA datasets. In addition, it reduces training time by 28.32% compared to its fully trained dense counterpart while maintaining accuracy, demonstrating clear efficiency gains. To further improve data efficiency, we propose a key-subset selection algorithm that selects and uses approximately 25% of MUSIC-AVQA v2.0 training data and retains 70–80% of full-data performance across models.

Dynamic and Generalizable Process Reward Modeling

Zhangyue Yin, Qiushi Sun, Zhiyuan Zeng, Qinyuan Cheng, Xipeng Qiu, Xuanjing Huang

Process Reward Models (PRMs) are crucial for guiding Large Language Models (LLMs) in complex scenarios by providing dense reward signals. However, existing PRMs primarily rely on heuristic approaches, which struggle with cross-domain generalization. While LLM-as-judge has been proposed to provide generalized rewards, current research has focused mainly on feedback results, overlooking the meaningful guidance embedded within the text. Additionally, static and coarse-grained evaluation criteria struggle to adapt to complex process supervision. To tackle these challenges, we propose Dynamic and Generalizable Process Reward Modeling (DG-PRM), which features a reward tree to capture and store fine-grained, multi-dimensional reward criteria. DG-PRM dynamically selects reward signals for step-wise reward scoring. To handle multifaceted reward signals, we pioneeringly adopt Pareto dominance estimation to identify discriminative positive and negative pairs. Experimental results show that DG-PRM achieves stunning performance on prevailing benchmarks, significantly boosting model performance across tasks with dense rewards. Further analysis reveals that DG-PRM adapts well to out-of-distribution scenarios, demonstrating exceptional generalizability.

Principled Understanding of Generalization for Generative Transformer Models in Arithmetic Reasoning Tasks

Xingcheng Xu, Zibo Zhao, Haipeng Zhang, Yanqing Yang

Transformer-based models excel in various tasks but their generalization capabilities, especially in arithmetic reasoning, remain incompletely understood. Arithmetic tasks provide a controlled framework to explore these capabilities, yet performance anomalies persist, such as inconsistent effectiveness in multiplication and erratic generalization in modular addition (e.g., modulo 100 vs. 101). This paper develops a unified theoretical framework for understanding the generalization behaviors of transformers in arithmetic tasks, focusing on length generalization. Through detailed analysis of addition, multiplication, and modular operations, we reveal that translation invariance in addition aligns with relative positional encoding for robust generalization, while base mismatch in modular operations disrupts this alignment. Experiments across GPT-family models validate our framework, confirming its ability to predict generalization behaviors. Our work highlights the importance of task structure and training data distribution for achieving data-efficient and structure-aware training, providing a systematic approach to understanding of length generalization in transformers.

Lost in the Context: Insufficient and Distracted Attention to Contexts in Preference Modeling

Shihan Dou, Jiayi Chen, Chenhao Huang, Feng Chen, Wei Chengzhi, Huiyuan Zheng, Shichun Liu, Yan Liu, Chenxiao Liu, Chao Xin, Lin Yan, Zongzhang Zhang, Tao Gui, Qi Zhang, Xuanjing Huang

In Reinforcement Learning from Human Feedback (RLHF), the reward model (RM) evaluates the response quality based on the given context and assigns a reward. It plays a crucial role in aligning RLHF with human preferences. Although the current RM training paradigm concatenates the context and response while amplifying the reward difference between good and bad response pairs, we demonstrate that the RM faces two significant issues: i) it often allocates only a small proportion of attention to the context, and ii) it frequently ignores segments of the context that are relevant for evaluating the response quality. These issues undermine the RM's effectiveness in modeling human preferences. To further address these challenges, we propose AttnRM, a novel optimization framework that enables the RM to concentrate on crucial segments of the context. Experimental results demonstrate that AttnRM significantly improves preference modeling by increasing attention to relevant information within the context. It also enhances the RM's generalizability and achieves better performance in aligning with human preferences.

Benchmarking Systematic Relational Reasoning with Large Language and Reasoning Models

Irtaza Khalid, Amir Masoud Nourollah, Steven Schockaert

Large Language Models (LLMs) have been found to struggle with systematic reasoning. Even on tasks where they appear to perform well, their performance often depends on shortcuts, rather than on genuine reasoning abilities, leading them to collapse on out-of-distribution (OOD) examples. Post-training strategies based on reinforcement learning and chain-of-thought prompting have recently been hailed as a step change. However, little is known about the potential of the resulting ``Large Reasoning Models" (LRMs) beyond maths and programming-based problem solving, where genuine OOD problems can be sparse. In this paper, we focus on tasks that require systematic relational composition for qualitative spatial and temporal reasoning. The setting allows fine control over problem difficulty to precisely measure OOD generalization. We find that, zero-shot LRMs generally outperform their LLM counterparts in single-path reasoning tasks but struggle in the multi-path setting. Whilst showing comparatively better results, fine-tuned LLMs are also not capable of multi-path generalization. We also provide evidence for the behavioral interpretation for this, i.e., that LRMs are shallow disjunctive reasoners.

Circuit Stability Characterizes Language Model Generalization

Alan Sun

Extensively evaluating the capabilities of (large) language models is difficult. Rapid development of state-of-the-art models induce benchmark saturation, while creating more challenging datasets is labor-intensive. Inspired by the recent developments in mechanistic interpretability, we introduce circuit stability as a new way to assess model performance. Circuit stability refers to a model's ability to apply a consistent reasoning process–its circuit–across various inputs. We mathematically formalize circuit stability and circuit equivalence. Then, through three case studies, we empirically show that circuit stability and the lack thereof can characterize and predict different aspects of generalization. Our proposed methods offer a step towards rigorously relating the generality of models to their interpretability.

IRIS: An Iterative and Integrated Framework for Verifiable Causal Discovery in the Absence of Tabular Data

Tao Feng, Lizhen Qu, Niket Tandon, Gholamreza Haffari

Causal discovery is fundamental to scientific research, yet traditional statistical algorithms face significant challenges, including expensive data collection, redundant computation for known relations, and unrealistic assumptions. While recent LLM-based methods excel at identifying commonly known causal relations, they fail to uncover novel relations. We introduce IRIS (Iterative Retrieval and Integrated System for Real-Time Causal Discovery), a novel framework that addresses these limitations. Starting with a set of initial variables, IRIS automatically collects relevant documents, extracts variables, and uncovers causal relations. Our hybrid causal discovery method combines statistical algorithms and LLM-based methods to discover known and novel causal relations. In addition to causal discovery on initial variables, the missing variable proposal component of IRIS identifies and incorporates missing variables to expand the causal graphs. Our approach enables real-time causal discovery from only a set of initial variables without requiring pre-existing datasets.

ExpeTrans: LLMs Are Experiential Transfer Learners

Jinglong Gao, Xiao Ding, Lingxiao Zou, Bibo Cai, Bing Qin, Ting Liu

Recent studies provide large language models (LLMs) with textual task-solving experiences via prompts to improve their performance. However, previous methods rely on substantial human labor or time to gather such experiences for each task, which is impractical given the growing variety of task types in user queries to LLMs. To address this issue, we design an autonomous experience transfer framework to explore whether LLMs can mimic human cognitive intelligence to autonomously transfer experience from existing source tasks to newly encountered target tasks. This not only allows the acquisition of experience without extensive costs of previous methods, but also offers a novel path for the generalization of LLMs. Experimental results on 13 datasets demonstrate that our framework effectively improves the performance of LLMs. Furthermore, we provide a detailed analysis of each module in the framework.

Towards Robust ESG Analysis Against Greenwashing Risks: Aspect-Action Analysis with Cross-Category Generalization

Keane Ong, Rui Mao, Deeksha varshney, Erik Cambria, Gianmarco Mengaldo

Sustainability reports are key for evaluating companies' environmental, social and governance (ESG) performance. To analyze these reports, NLP approaches can efficiently extract ESG insights at scale. However, even the most advanced NLP methods lack robustness against ESG content that is greenwashed – i.e. sustainability claims that are misleading, exaggerated, and fabricated. Accordingly, existing NLP approaches often extract insights that reflect misleading or exaggerated sustainability claims rather than objective ESG performance. To tackle this issue, we introduce A3CG - **A**spect-**A**ction **A**nalysis with Cross-**C**ategory **G**eneralization, as a novel dataset to improve the robustness of ESG analysis amid the prevalence of greenwashing. By explicitly linking sustainability aspects with their associated actions, A3CG facilitates a more fine-grained and transparent evaluation of sustainability claims, ensuring that insights are grounded in verifiable actions rather than vague or misleading rhetoric. Additionally, A3CG emphasizes cross-category generalization. This ensures robust model performance in aspect-action analysis even when companies change their reports to selectively favor certain sustainability areas. Through experiments on A3CG, we analyze state-of-the-art supervised models and LLMs, uncovering their limitations and outlining key directions for future research.

MolRAG: Unlocking the Power of Large Language Models for Molecular Property Prediction

Ziting Xian, Jiawei Gu, Lingbo Li, Eran Segal, Shangsong Liang

Recent LLMs exhibit limited effectiveness on molecular property prediction task due to the semantic gap between molecular representations and natural language, as well as the lack of domain-specific knowledge. To address these challenges, we propose MolRAG, a Retrieval-Augmented Generation framework integrating Chain-of-Thought reasoning for molecular property prediction. MolRAG operates by retrieving structurally analogous molecules as contextual references to guide stepwise knowledge reasoning through chemical structure-property relationships. This dual mechanism synergizes molecular similarity analysis with structured inference, while generating human-interpretable rationales grounded in domain knowledge. Experimental results show MolRAG outperforms pre-trained LLMs on four datasets, and even matches supervised methods, achieving performance gains of 1.1%–45.7% over direct prediction approaches, demonstrating versatile effectiveness. Our code is available at https://github.com/AcaciaSin/MolRAG.

How to Mitigate Overfitting in Weak-to-strong Generalization?

Junhao Shi, Qinyuan Cheng, Zhaoye Fei, Yining Zheng, Qipeng Guo, Xipeng Qiu

Aligning powerful AI models on tasks that surpass human evaluation capabilities is the central problem of **superalignment**. To address this problem, weak-to-strong generalization aims to elicit the capabilities of strong models through weak supervisors and ensure that the behavior of strong models aligns with the intentions of weak supervisors without unsafe behaviors such as deception. Although weak-to-strong generalization exhibiting certain generalization capabilities, strong models exhibit significant overfitting in weak-to-strong generalization: Due to the strong fit ability of strong models, erroneous labels from weak supervisors may lead to overfitting in strong models. In addition, simply filtering out incorrect labels may lead to a degeneration in question quality, resulting in a weak generalization ability of strong models on hard questions. To mitigate overfitting in weak-to-strong generalization, we propose a two-stage framework that simultaneously improves the quality of supervision signals and the quality of input questions. Experimental results in three series of large language models and two mathematical benchmarks demonstrate that our framework significantly improves PGR (Performance Gap Recovered) compared to naive weak-to-strong generalization, even achieving up to 100% PGR on some models.

Acoustic Individual Identification of White-Faced Capuchin Monkeys Using Joint Multi-Species Embeddings

Álvaro Vega-Hidalgo, Artem Abzaliev, Thore Bergman, Rada Mihalcea

Acoustic individual identification of wild animals is an essential task for understanding animal vocalizations within their social contexts, and for facilitating conservation and wildlife monitoring efforts. However, most of the work in this space relies on human efforts, as the development of methods for automatic individual identification is hindered by the lack of data. In this paper, we explore cross-species pre-training to address the task of individual classification in white-faced capuchin monkeys. Using acoustic embeddings from birds and humans, we find that they can be effectively used to identify the calls from individual monkeys. Moreover, we find that joint multi-species representations can lead to further improvements over the use of one representation at a time. Our work demonstrates the potential of cross-species data transfer and multi-species representations, as strategies to address tasks on species with very limited data.

Can LLMs Help Uncover Insights about LLMs? A Large-Scale, Evolving Literature Analysis of Frontier LLMs

Jungsoo Park, Junmo Kang, Gabriel Stanovsky, Alan Ritter

The surge of LLM studies makes synthesizing their findings challenging. Analysis of experimental results from literature can uncover important trends across studies, but the time-consuming nature of manual data extraction limits its use. Our study presents a semi-automated approach for literature analysis that accelerates data extraction using LLMs. It automatically identifies relevant arXiv papers, extracts experimental results and related attributes, and organizes them into a structured dataset, LLMEvalDB. We then conduct an automated literature analysis of frontier LLMs, reducing the effort of paper surveying and data extraction by more than 93% compared to manual approaches. We validate LLMEvalDB by showing that it reproduces key findings from a recent manual analysis of Chain-of-Thought (CoT) reasoning and also uncovers new insights that go beyond it, showing, for example, that in-context examples benefit coding \& multimodal tasks but offer limited gains in math reasoning tasks compared to zero-shot CoT. Our automatically updatable dataset enables continuous tracking of target models by extracting evaluation studies as new data becomes available. Through LLMEvalDB and empirical analysis, we provide insights into LLMs while facilitating ongoing literature analyses of their behavior.

GenKnowSub: Improving Modularity and Reusability of LLMs through General Knowledge Subtraction

Mohammadtaha Bagherifard, Sahar Rajabi, Ali Edalat, Yadollah Yaghoobzadeh

Large language models (LLMs) often struggle with zero-shot generalization, and several modular approaches have been proposed to address this challenge. Yet, we hypothesize that a key limitation remains: the entanglement of general knowledge and task-specific adaptations. To overcome this, we propose a modular framework that disentangles these components by constructing a library of task-specific LoRA modules alongside a general-domain LoRA. By subtracting this general knowledge component from each task-specific module, we obtain residual modules that focus more exclusively on task-relevant information. We call this approach general knowledge subtraction or GenKnowSub. Leveraging the refined task-specific modules and the Arrow routing algorithm, we dynamically select and combine modules for new inputs without additional training. Our studies on the Phi-3 model and standard Arrow as baselines reveal that using general knowledge LoRAs derived from diverse languages, including English, French, and German, yields consistent performance gains in both monolingual and cross-lingual settings across a wide set of benchmarks. Further experiments on Phi-2 reveal how GenKnowSub generalizes to a weaker LLM.

On Generalization across Measurement Systems: LLMs Entail More Test-Time Compute for Underrepresented Cultures

Minh Duc Bui, Kyung eun Park, Goran Glavaš, Fabian David Schmidt, Katharina von der Wense

Measurement systems (e.g., currencies) differ across cultures, but the conversions between them are well defined so that humans can state using any measurement system of their choice. Being available to users from diverse cultural backgrounds, Large Language Models (LLMs) should also be able to provide accurate information irrespective of the measurement system at hand. Using newly compiled datasets we test if this is truly the case for seven open-source LLMs, addressing three key research questions: (RQ1) What is the default system used by LLMs for each type of measurement? (RQ2) Do LLMs' answers and their accuracy vary across different measurement systems? (RQ3) Can LLMs mitigate potential challenges w.r.t. underrepresented systems via reasoning? Our findings show that LLMs default to the measurement system predominantly used in the data. Additionally, we observe considerable instability and variance in performance across different measurement systems. While this instability can in part be mitigated by employing reasoning methods such as chain-of-thought (CoT), this implies longer responses and thereby significantly increases test-time compute (and inference costs), marginalizing users from cultural backgrounds that use underrepresented measurement systems.

Can LLMs Understand Unvoiced Speech? Exploring EMG-to-Text Conversion with LLMs

Payal Mohapatra, Akash Pandey, Xiaoyuan Zhang, Qi Zhu

Unvoiced electromyography (EMG) is an effective communication tool for individuals unable to produce vocal speech. However, most prior methods rely on paired voiced and unvoiced EMG signals, along with speech data, for unvoiced EMG-to-text conversion, which is not practical for these individuals. Given the rise of large language models (LLMs) in speech recognition, we explore their potential to understand unvoiced speech. To this end, we address the challenge of \emph{learning from unvoiced EMG alone} and propose a novel EMG adaptor module that maps EMG features to an LLM's input space, achieving an average word error rate of 0.49 on a closed-vocabulary unvoiced EMG-to-text task. Even with a conservative data availability of just six minutes, our approach improves performance over specialized models by nearly 20%. While LLMs have been shown to be extendable to new language modalities—such as audio—understanding articulatory biosignals, like unvoiced EMG, is more challenging. This work takes a crucial first step toward enabling LLMs to comprehend unvoiced speech using surface EMG.

"What do you call a dog that is incontrovertibly true? Dogma": Testing LLM Generalization through Humor

Alessio Cocchieri, Luca Ragazzi, Paolo Italiani, Giuseppe Tagliavini, Gianluca Moro

Humor, requiring creativity and contextual understanding, is a hallmark of human intelligence, showcasing adaptability across linguistic scenarios. While recent advances in large language models (LLMs) demonstrate strong reasoning on various benchmarks, it remains unclear whether they truly adapt to new tasks like humans (i.e., generalize) or merely replicate memorized content. To explore this, we introduce Phunny, a new humor-based question-answering benchmark designed to assess LLMs' reasoning through carefully crafted puns. Our dataset is manually curated to ensure novelty and minimize data contamination, providing a robust evaluation of LLMs' linguistic comprehension. Experiments on pun comprehension, resolution, and generation reveal that most LLMs struggle with generalization, even on simple tasks, consistently underperforming the human baseline. Additionally, our detailed error analysis provides valuable insights to guide future research.

Exploring Explanations Improves the Robustness of In-Context Learning

Ukyo Honda, Tatsushi Oka

In-context learning (ICL) has emerged as a successful paradigm for leveraging large language models (LLMs). However, it often struggles to generalize beyond the distribution of the provided demonstrations. A recent advancement in enhancing robustness is ICL with explanations (X-ICL), which improves prediction reliability by guiding LLMs to understand and articulate the reasoning behind correct labels. Building on this approach, we introduce an advanced framework that extends X-ICL by systematically exploring explanations for all possible labels (X^2-ICL), thereby enabling more comprehensive and robust decision-making. Experimental results on multiple natural language understanding datasets validate the effectiveness of X^2-ICL, demonstrating significantly improved robustness to out-of-distribution data compared to the existing ICL approaches.

Unravelling the Logic: Investigating the Generalisation of Transformers in Numerical Satisfiability Problems

Tharindu Madusanka, Marco Valentino, Iqra Zahid, Ian Pratt-Hartmann, Riza Batista-Navarro

Transformer models have achieved remarkable performance in many formal reasoning tasks. Nonetheless, the extent of their comprehension pertaining to logical semantics and rules of inference remains somewhat uncertain. Evaluating such understanding necessitates a rigorous examination of these models' generalisation capacity to out-of-distribution data. In this study, we probe the generalisation prowess of Transformer models with respect to the hitherto unexplored domain of numerical satisfiability problems. Our investigation reveals that Transformers exhibit minimal scale and noise invariance, alongside limited vocabulary and number invariance. However, even when Transformer models experience a notable decline in performance on out-of-distribution test sets, they often still surpass the random baseline by a considerable margin.

Dedicated Feedback and Edit Models Empower Inference-Time Scaling for Open-Ended General-Domain Tasks

Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Daniel Egert, Ellie Evans, Hoo-Chang Shin, Felipe Soares, Yi Dong, Oleksii Kuchaiev

Inference-Time Scaling has been critical to the success of recent models such as OpenAI o1 and DeepSeek R1. However, many techniques used to train models for inference-time scaling require tasks to have answers that can be verified, limiting their application to domains such as math, coding and logical reasoning. We take inspiration from how humans make first attempts, ask for detailed feedback from others and make improvements based on such feedback across a wide spectrum of open-ended endeavors. To this end, we collect HelpSteer3 data to train dedicated Feedback and Edit Models that are capable of performing inference-time scaling for open-ended general-domain tasks. In our setup, one model generates an initial response, which are given feedback by a second model, that are then used by a third model to edit the response. We show that performance on Arena Hard, a benchmark strongly predictive of Chatbot Arena Elo can be boosted by scaling the number of initial response drafts, effective feedback and edited responses. When scaled optimally, our setup based on 70B models from the Llama 3 family can reach SoTA performance on Arena Hard at 92.7 as of 5 Mar 2025, surpassing OpenAI o1-preview-2024-09-12 with 90.4 and DeepSeek R1 with 92.3.

BIG-Bench Extra Hard

Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, Lalit K Jain, Virginia Aglietti, Disha Jindal, Peter Chen, Nishanth Dikkala, Gladys Tyen, Xin Liu, Uri Shalit, Silvia Chiappa, Kate Olszewska, Yi Tay, Vinh Q. Tran, Quoc V Le, Orhan Firat

Current benchmarks for large language model (LLM) reasoning predominantly focus on mathematical and coding abilities, leaving a gap in evaluating broader reasoning proficiencies. One particular exception is the BIG-Bench dataset, which has served as a crucial benchmark for evaluating the general reasoning capabilities of LLMs, thanks to its diverse set of challenging tasks that allowed for a comprehensive assessment of general reasoning across various skills within a unified framework. However, recent advances in LLMs have led to saturation on BIG-Bench, and its harder version BIG-Bench Hard (BBH). State-of-the-art models achieve near-perfect scores on many tasks in BBH, thus diminishing its utility. To address this limitation, we introduce BIG-Bench Extra Hard (BBEH), a new benchmark designed to push the boundaries of LLM reasoning evaluation. BBEH replaces each task in BBH with a novel task that probes a similar reasoning capability but exhibits significantly increased difficulty. We evaluate various general-purpose and reasoning-specialized models on BBEH and observe an accuracy of 23.9% for the best general-purpose model and 54.2% for the best reasoning-specialized model, indicating substantial room for improvement and highlighting the ongoing challenge of achieving robust general reasoning in LLMs. We release BBEH publicly at: https: //github.com/google-deepmind/bbeh.

Dialectal Coverage And Generalization in Arabic Speech Recognition

Amirbek Djanibekov, Hawau Olamide Toyin, Raghad Alshalan, Abdullah Alatir, Hanan Aldarmaki

Developing robust automatic speech recognition (ASR) systems for Arabic requires effective strategies to manage its diversity. Existing ASR systems mainly cover the modern standard Arabic (MSA) variety and few high-resource dialects, but fall short in coverage and generalization across the multitude of spoken variants. Code-switching with English and French is also common in different regions of the Arab world, which challenges the performance of monolingual Arabic models. In this work, we introduce a suite of ASR models optimized to effectively recognize multiple variants of spoken Arabic, including MSA, various dialects, and code-switching. We provide open-source pre-trained models that cover data from 17 Arabic-speaking countries, and fine-tuned MSA and dialectal ASR models that include at least 11 variants, as well as multi-lingual ASR models covering embedded languages in code-switched utterances. We evaluate ASR performance across these spoken varieties and demonstrate both coverage and performance gains compared to prior models.

Stochastic Chameleons: Irrelevant Context Hallucinations Reveal Class-Based (Mis)Generalization in LLMs

Ziling Cheng, Meng Cao, Marc-Antoine Rondeau, Jackie CK Cheung

The widespread success of LLMs on NLP benchmarks has been accompanied by concerns that LLMs function primarily as stochastic parrots that reproduce texts similar to what they saw during pre-training, often erroneously. But what is the nature of their errors, and do these errors exhibit any regularities? In this work, we examine irrelevant context hallucinations, in which models integrate misleading contextual cues into their predictions. Through behavioral analysis, we show that these errors result from a structured yet flawed mechanism that we term _class-based (mis)generalization_, in which models combine abstract class cues with features extracted from the query or context to derive answers. Furthermore, mechanistic interpretability experiments on Llama-3, Mistral, and Pythia across 39 factual recall relation types reveal that this behavior is reflected in the model's internal computations: (i) abstract class representations are constructed in lower layers before being refined into specific answers in higher layers, (ii) feature selection is governed by two competing circuits --- one prioritizing direct query-based reasoning, the other incorporating contextual cues --- whose relative influences determine the final output. Our findings provide a more nuanced perspective on the stochastic parrot argument: through form-based training, LLMs can exhibit generalization leveraging abstractions, albeit in unreliable ways based on contextual cues — what we term _stochastic chameleons_.

Session 10: Special Theme: Generalization of NLP Models (Tuesday 29 July, 16:00-17:30, Hall 5X)

DECAF: A Dynamically Extensible Corpus Analysis Framework

Max Müller-Eberstein, Rob van der Goot, Anna Rogers

The study of generalization in Language Models (LMs) requires controlled experiments that can precisely measure complex linguistic variations between training and testing datasets. We introduce DECAF, a framework that enables the analysis and filtering of linguistically-annotated datasets down to the character level. Rather than creating new resources for each experiment, DECAF starts from datasets with existing linguistic annotations, and leverages them to analyze, filter, and generate highly controlled and reproducible experimental settings targeting specific research questions. We demonstrate DECAF's functionality by adding full morphology and dependency annotations to the 131M-token BabyLM corpus and indexing the resulting 1.1B annotations to analyze its internal domain variance, and to create a controlled training data curriculum for a small-scale gender bias study. We release DECAF as a fully open-source Python library, along with the parsed and indexed version of BabyLM, as resources for future generalization research.

Session 10: Student Research Workshop (Tuesday 29 July, 16:00-17:30, Hall 4/5)

Learning and Enforcing Context-Sensitive Control for LLMs

Mohammad Albinhassan, Pranava Madhyastha, Mark Law, Alessandra Russo

Controlling the output of Large Language Models (LLMs) through context-sensitive constraints has emerged as a promising approach to overcome the limitations of Context-Free Grammars (CFGs) in guaranteeing generation validity. However, such constraints typically require manual specification---a significant barrier demanding specialized expertise. We introduce a framework that automatically learns context-sensitive constraints from LLM interactions through a two-phase process: syntactic exploration to gather diverse outputs for constraint learning, followed by constraint exploitation to enforce these learned rules during generation. Experiments demonstrate that our method enables even small LLMs (1B parameters) to learn and generate with perfect constraint adherence, outperforming larger counterparts and state-of-the-art reasoning models. This work represents the first integration of context-sensitive grammar learning with LLM generation, eliminating manual specification while maintaining generation validity.

Unwrapping Circularity: Can Transformers Learn Languages with Circular Schemes?

Xiutian Zhao, Aulia Rafi, Siying Chen, Xiulin Yang

The success of Transformer-based language models in NLP has sparked debate about their ability to simulate human language learning. Chomsky contends that these models indiscriminately acquire both natural and "impossible" languages. While recent studies have challenged this claim, the capacity of Transformers to handle unconventional linguistic structures remains underexplored. Inspired by natural and speculative languages with circular structural properties, this study examines the ability of GPT-2 to learn languages featuring circular schemes. We synthesize such circular languages by mapping original sequences onto textual circles and then relinearize them using parametric, mathematically invertible procedures that "unwrap" the circles into linear sequences. We train GPT-2 models on these relinearized corpora and assess the impact of linearization parameters by tracking structural distortion and measuring perplexity. Interestingly, high levels of distortion relative to the original structures do not necessarily correspond to increased perplexity, suggesting that GPT-2 is relatively insensitive to global token order during language acquisition. Instead, preserving local context during linearization plays a more critical role in model learning. Further analysis using surprisal differences reveals that positional shifts pose greater challenges to the model than changes in stride or direction, underscoring the nuanced effects of linearization strategies. These findings offer new insights into the inductive biases of Transformer-based models in acquiring unconventional linguistic structures.

When Will the Tokens End? Graph-Based Forecasting for LLMs Output Length

Grzegorz Piotrowski, Mateusz Bystroski, Mikoaj Hoysz, Jakub Binkowski, Grzegorz Chodak, Tomasz Jan Kajdanowicz

Large Language Models (LLMs) are typically trained to predict the next token in a sequence. However, their internal representations often encode signals that go beyond immediate next-token prediction. In this work, we investigate whether these hidden states also carry information about the remaining length of the generated output—an implicit form of foresight \cite{pal-etal-2023-future}. We formulate this as a regression problem where, at generation step t, the target is the number of remaining tokens y_t = T - t, with T as the total output length. We propose two approaches: (1) an aggregation-based model that combines hidden states from multiple transformer layers \ell \in \{8, \dots, 15\} using element-wise operations such as mean or sum, and (2) a \textit{Layerwise Graph Regressor} that treats layerwise hidden states as nodes in a fully connected graph and applies a Graph Neural Network (GNN) to predict y_t. Both models operate on frozen LLM embeddings without requiring end-to-end fine-tuning. Accurately estimating remaining output length has both theoretical and practical implications. From an interpretability standpoint, it suggests that LLMs internally track their generation progress. From a systems perspective, it enables optimizations such as output-length-aware scheduling \cite{shahout2024dontstopnowembedding}. Our graph-based model achieves state-of-the-art performance on the Alpaca dataset using LLaMA-3-8B-Instruct, reducing normalized mean absolute error (NMAE) by over 50\% in short-output scenarios.

Only for the Unseen Languages, Say the Llamas: On the Efficacy of Language Adapters for Cross-lingual Transfer in English-centric LLMs

Julian Schlenker, Jenny Kunz, Tatiana Anikina, Gnter Neumann, Simon Ostermann

Most state-of-the-art large language models (LLMs) are trained mainly on English data, limiting their effectiveness on non-English, especially low-resource, languages. This study investigates whether language adapters can facilitate cross-lingual transfer in English-centric LLMs. We train language adapters for 13 languages using Llama 2 (7B) and Llama 3.1 (8B) as base models, and evaluate their effectiveness on two downstream tasks (MLQA and SIB-200) using either task adapters or in-context learning. Our results reveal that language adapters improve performance for languages not seen during pretraining, but provide negligible benefit for seen languages. These findings highlight the limitations of language adapters as a general solution for multilingual adaptation in English-centric LLMs.

HyILR: Hyperbolic Instance-Specific Local Relationships for Hierarchical Text Classification

Ashish Kumar, Durga Toshniwal

Recent approaches to Hierarchical Text Classification (HTC) rely on capturing the global label hierarchy, which contains static and often redundant relationships. Instead, the hierarchical relationships within the instance-specific set of positive labels are more important, as they focus on the relevant parts of the hierarchy. These localized relationships can be modeled as a semantic alignment between the text and its positive labels within the embedding space. However, without explicitly encoding the global hierarchy, achieving this alignment directly in Euclidean space is challenging, as its flat geometry does not naturally support hierarchical relationships. To address this, we propose Hyperbolic Instance-Specific Local Relationships (HyILR), which models instance-specific relationships using the Lorentz model of hyperbolic space. Text and label features are projected into hyperbolic space, where a contrastive loss aligns text with its labels. This loss is guided by a hierarchy-aware negative sampling strategy, ensuring the selection of structurally and semantically relevant negatives. By leveraging hyperbolic geometry for this alignment, our approach inherently captures hierarchical relationships and eliminates the need for global hierarchy encoding. Experimental results on four benchmark datasets validate the superior performance of HyILR over baseline methods.

Pragmatic Perspective on Assessing Implicit Meaning Interpretation in Sentiment Analysis Models

Rashid Mustafin

Drawing on pragmatic theories of implicature by Grice (1975) and Levinson (1983), according to which speakers often convey more than it is explicitly said, the paper argues that interpreting texts with implicit meaning correctly is essential for precise sentiment analysis. To illustrate the challenges in computational interpretation of implicatures, the study introduces a series of illustrative micro-experiments with the use of four transformer models fine-tuned for sentiment analysis. In these micro-experiments, the models classified sentences specifically designed to expose difficulties in handling implicit meaning. The study demonstrates that contrasting qualitative pragmatic analysis with the models' tendency to focus on formal linguistic markers can reveal the limitations of supervised machine learning methods in detecting implicit sentiments.

Foundations of PEERS: Assessing LLM Role Performance in Educational Simulations

Jasper Meynard Arana, Kristine Ann M. Carandang, Ethan Robert Casin, Christian Alis, Daniel Stanley Tan, Erika Fille Legara, Christopher Monterola

In education, peer instruction (PI) is widely recognized as an effective active learning strategy. However, real-world evaluations of PI are often limited by logistical constraints and variability in classroom settings. This paper introduces PEERS (Peer Enhanced Educational Realistic Simulation), a simulation framework that integrates Agent-Based Modeling (ABM), Large Language Models (LLMs), and Bayesian Knowledge Tracing (BKT) to emulate student learning dynamics. As an initial step, this study focuses on evaluating whether LLM-powered agents can effectively assume the roles of teachers and students within the simulation. Human evaluations and topic-based metrics show that LLMs can generate role-consistent and contextually appropriate classroom dialogues. These results serve as a foundational milestone toward building realistic, AI-driven educational simulations. Future work will include simulating the complete PEERS framework and validating its accuracy through actual classroom-based PI sessions. This research aims to contribute a scalable, cost-effective methodology for studying instructional strategies in controlled yet realistic environments.

The Role of Exploration Modules in Small Language Models for Knowledge Graph Question Answering

Yi-Jie Cheng, Oscar Chew, Yun-Nung Chen

Integrating knowledge graphs (KGs) into the reasoning processes of large language models (LLMs) has emerged as a promising approach to mitigate hallucination. However, existing work in this area often relies on proprietary or extremely large models, limiting accessibility and scalability. In this study, we investigate the capabilities of existing integration methods for small language models (SLMs) in KG-based question answering and observe that their performance is often constrained by their limited ability to traverse and reason over knowledge graphs. To address this limitation, we propose leveraging simple and efficient exploration modules to handle knowledge graph traversal in place of the language model itself. Experiment results demonstrate that these lightweight modules effectively improve the performance of small language models on knowledge graph question answering tasks. Our code will be available on Github.

Building Japanese Creativity Benchmarks and Applying them to Enhance LLM Creativity

So Fukuda, Hayato Ogawa, Kaito Horio, Daisuke Kawahara, Tomohide Shibata

To evaluate the creativity of large language models (LLMs) in Japanese, we construct three benchmarks: Japanese Creativity Questions (JCQ), Divergent Association Task (DAT), and Story Alteration Task (SAT). JCQ comprehensively evaluates creativity using LLMs. Meanwhile, DAT and SAT measure specific aspects of creative ability using embeddings. We also analyze correlations between JCQ and DAT, JCQ and SAT, and DAT and SAT. While JCQ provides comprehensive evaluation, it is relatively time and resource intensive. In contrast, DAT and SAT offer lower comprehensiveness but enable quick, low-cost assessment. Additionally, we investigate whether training with DAT contributes to enhancing LLM creativity.

Towards Robust Sentiment Analysis of Temporally-Sensitive Policy-Related Online Text

Charles Alba, Benjamin C Warner, Akshar Saxena, Jiaxin Huang, Ruopeng An

Sentiment analysis in policy-related studies typically involves annotating a subset of data to fine-tune a pre-trained model, which is subsequently used to classify sentiments in the remaining unlabeled texts, enabling policy researchers to analyze sentiments in novel policy contexts under resource constraints. We argue that existing methods fail to adequately capture the temporal volatility inherent in policy-related sentiments, which are subject to external shocks and evolving discourse of opinions. We propose methods accounting for the temporal dynamics of policy-related texts. Specifically, we propose leveraging continuous time-series clustering to select data points for annotation based on temporal trends and subsequently apply model merging techniques -- each fine-tuned separately on data from distinct time intervals. Our results indicate that continuous time-series clustering followed by fine-tuning a single unified model achieves superior performance, outperforming existing methods by an average F1-score of 2.71\%. This suggests that language models can generalize to temporally sensitive texts when provided with temporally representative samples. Nevertheless, merging multiple time-specific models - particularly via greedy soup and TIES - achieves competitive performance, suggesting practical applications in dynamically evolving policy scenarios.

Is Partial Linguistic Information Sufficient for Discourse Connective Disambiguation? A Case Study of Concession

Takuma Sato, Ai Kubota, Koji Mineshima

Discourse relations are sometimes explicitly conveyed by specific connectives. However, some connectives can signal multiple discourse relations; in such cases, disambiguation is necessary to determine which relation is intended. This task is known as *discourse connective disambiguation* (Pitler and Nenkova, 2009), and particular attention is often given to connectives that can convey both *concession* and other relations (e.g., *synchronous*). In this study, we conducted experiments to analyze which linguistic features play an important role in the disambiguation of polysemous connectives in Japanese. A neural language model (BERT) was fine-tuned using inputs from which specific linguistic features (e.g., word order, specific lexicon, etc.) had been removed. We analyzed which linguistic features affect disambiguation by comparing the model's performance. Our results show that even after performing drastic removal, such as deleting one of the two arguments that constitute the discourse relation, the model's performance remained relatively robust. However, the removal of certain lexical items or words belonging to specific lexical categories significantly degraded disambiguation performance, highlighting their importance in identifying the intended discourse relation.

Semantic Frame Induction from a Real-World Corpus

Shogo Tsujimoto, Kosuke Yamada, Ryohei Sasano

Recent studies on semantic frame induction have demonstrated that the emergence of pre-trained language models (PLMs) has led to more accurate results. However, most existing studies evaluate the performance using frame resources such as FrameNet, which may not accurately reflect real-world language usage. In this study, we conduct semantic frame induction using the Colossal Clean Crawled Corpus (C4) and assess the applicability of existing frame induction methods to real-world data. Our experimental results demonstrate that existing frame induction methods are effective on real-world data and that frames corresponding to novel concepts can be induced.

Lost and Found: Computational Quality Assurance of Crowdsourced Knowledge on Morphological Defectivity in Wiktionary

Jonathan Sakunkoo, Annabella Sakunkoo

Morphological defectivity is an intriguing and understudied phenomenon in linguistics. Addressing defectivity, where expected inflectional forms are absent, is essential for improving the accuracy of NLP tools in morphologically rich languages. However, traditional linguistic resources often lack coverage of morphological gaps as such knowledge requires significant human expertise and effort to document and verify. For scarce linguistic phenomena in under-explored languages, Wikipedia and Wiktionary often serve as among the few accessible resources. Despite their extensive reach, their reliability has been a subject of controversy. This study customizes a novel neural morphological analyzer to annotate Latin and Italian corpora. Using the massive annotated data, crowd-sourced lists of defective verbs compiled from Wiktionary are validated computationally. Our results indicate that while Wiktionary provides a highly reliable account of Italian morphological gaps, 7% of Latin lemmata listed as defective show strong corpus evidence of being non-defective. This discrepancy highlights potential limitations of crowd-sourced wikis as definitive sources of linguistic knowledge, particularly for less-studied phenomena and languages, despite their value as resources for rare linguistic features. By providing scalable tools and methods for quality assurance of crowd-sourced data, this work advances computational morphology and expands linguistic knowledge of defectivity in non-English, morphologically rich languages.

From Directions to Cones: Multidimensional Representations of Propositional Facts in LLMs

Stanley Yu, Vaidehi Bulusu, Oscar S. Yasunaga, Clayton Lau, Cole Blondin, Vasu Sharma, Kevin Zhu, Sean O'Brien

Large Language Models (LLMs) exhibit strong conversational abilities but often generate falsehoods. Prior work suggests that the truthfulness of simple propositions can be represented as a single linear direction in a model's internal activations, but this may not fully capture its underlying geometry. In this work, we extend the concept cone framework—recently introduced for modeling refusal—to the domain of truth. We identify multi-dimensional cones that causally mediate truth-related behavior across multiple LLM families. Our results are supported by three lines of evidence: (i) causal interventions reliably flip model responses to factual statements; (ii) learned cones generalize across model architectures; and (iii) cone-based interventions preserve unrelated model behavior. These findings reveal the richer, multidirectional structure governing simple true/false propositions in LLMs and highlight concept cones as a promising tool for probing abstract behaviors.

Auto-TA: Towards Scalable Automated Thematic Analysis (TA) via Multi-Agent Large Language Models with Reinforcement Learning

Seungjun Yi, Joakim Nguyen, Huimin Xu, Terence Lim, Andrew Well, Mia Markey, Ying Ding

Congenital heart disease (CHD) presents complex, lifelong challenges often underrepresented in traditional clinical metrics. While unstructured narratives offer rich insights into patient and caregiver experiences, manual thematic analysis (TA) remains labor-intensive and unscalable. We propose a fully automated large language model (LLM) pipeline that performs end-to-end TA on clinical narratives which eliminates the need for manual coding or full transcript review. Our system employs a novel multi-agent framework, where specialized LLM agents assume roles to enhance theme quality and alignment with human analysis. To further improve thematic relevance, we optionally integrate reinforcement learning from human feedback (RLHF). This supports scalable, patient-centered analysis of large qualitative datasets and allows LLMs to be fine-tuned for specific clinical contexts.

Improving Explainability of Sentence-level Metrics via Edit-level Attribution for Grammatical Error Correction

Takumi Goto, Justin Vasselli, Taro Watanabe

Various evaluation metrics have been proposed for Grammatical Error Correction (GEC), but many, particularly reference-free metrics, lack explainability. This lack of explainability hinders researchers from analyzing the strengths and weaknesses of GEC models and limits the ability to provide detailed feedback for users. To address this issue, we propose attributing sentence-level scores to individual edits, providing insight into how specific corrections contribute to the overall performance. For the attribution method, we use Shapley values, from cooperative game theory, to compute the contribution of each edit. Experiments with existing sentence-level metrics demonstrate high consistency across different edit granularities and show approximately 70% alignment with human evaluations. In addition, we analyze biases in the metrics based on the attribution results, revealing trends such as the tendency to ignore orthographic edits.

Causal Language Control in Multilingual Transformers via Sparse Feature Steering

Cheng-Ting Chou, George Liu, Jessica Sun, Cole Blondin, Kevin Zhu, Vasu Sharma, Sean O'Brien

Deterministically controlling the target generation language of large multilingual language models (LLMs) remains a fundamental challenge, particularly in zero-shot settings where neither explicit language prompts nor fine-tuning are available. In this work, we investigate whether sparse autoencoder (SAE) features, previously shown to correlate with interpretable model behaviors, can be leveraged to steer the generated language of LLMs during inference. Leveraging pretrained SAEs on the residual streams of Gemma-2B and Gemma-9B, we identify features whose activations differ most significantly between English and four target languages: Chinese, Japanese, Spanish, and French. By modifying just a single SAE feature at one transformer layer, we achieve controlled language shifts with up to 90% success, as measured by FastText language classification, while preserving semantic fidelity according to LaBSE (Language-Agnostic BERT Sentence Embedding) similarity. Our analysis reveals that language steering is most effective in mid-to-late transformer layers and is amplified by specific attention heads disproportionately associated with language-sensitive SAE features. These results demonstrate the promise of sparse feature steering as a lightweight and interpretable mechanism for controllable multilingual generation.

From One-Fit-All to Perspective Aware Models: A Thesis Proposal

Leixin Zhang

Variation in human perspectives has drawn increasing attention in natural language processing. The widespread human annotation disagreement challenges the traditional paradigm of a single "ground truth" and raises concerns about the limitations of conventional label aggregation methods and the uniform models built upon them, which often discard minority opinions and obscure valuable individual perspectives. This thesis proposal investigates three core dimensions of perspective-oriented research: (1) annotation formats that better capture the nuance and uncertainty of individual judgments; (2) modeling approaches that leverage socio-demographic features to improve prediction for underrepresented or minority viewpoints; and (3) personalized generation that tailor outputs to individual users' preferences and communicative styles. Through this work, we aim to advance methods that more faithfully reflect the diversity of human interpretation, enhancing both inclusiveness and fairness in language technologies.

Semantic Convergence: Investigating Shared Representations Across Scaled LLMs

Andrew Rufail, Sanjana Rathore, Daniel Son, Adrian Simon, Soham Dave, Daniel Zhang, Cole Blondin, Sean O'Brien, Kevin Zhu

We investigate feature universality in Gemma-2 language models (Gemma-2-2B \& Gemma-2-9B), asking whether models with a fourfold difference in scale still converge on comparable internal concepts. Using the sparse autoencoder (SAE) dictionary learning pipeline, we used pretrained SAEs on each model's residual-stream activations, aligned the resulting monosemantic features via activation correlation, and compared the matched feature spaces with metrics such as SVCCA and RSA. Middle layers yield the strongest overlap, indicating that this is where both models most similarly represent concepts, while early and late layers show much less similarity. Preliminary experiments extending the analysis from single tokens to multi-token subspaces show that semantically similar subspaces tend to interact similarly with LLMs. These results offer further evidence that large language models carve the world into broadly similar, interpretable features despite size differences, reinforcing universality as a foundation for cross-model interpretability.

Controlling Language Confusion in Multilingual LLMs

Nahyun Lee, Yeongseo Woo, Hyunwoo Ko, Guijin Son

Large language models often suffer from language confusion, a phenomenon where responses are partially or entirely generated in unintended languages. This can critically impact user experience in low-resource settings. We hypothesize that conventional supervised fine-tuning exacerbates this issue because the softmax objective focuses probability mass only on the single correct token but does not explicitly penalize cross-lingual mixing. Interestingly, by observing loss trajectories during the pretraining phase, we observe that models fail to learn to distinguish between monolingual and language-confused text. Additionally, we find that ORPO, which adds penalties for unwanted output styles to standard SFT, effectively suppresses language-confused generations even at high decoding temperatures without degrading overall model performance. Our findings suggest that incorporating appropriate penalty terms can mitigate language confusion in low-resource settings with limited data.

Grammatical Error Correction via Sequence Tagging for Russian

Regina Nasyrova, Alexey Sorokin

We introduce a modified sequence tagging architecture, proposed in (Omelianchuk et al., 2020), for the Grammatical Error Correction of the Russian language. We propose language-specific operation set and preprocessing algorithm as well as a classification scheme which makes distinct predictions for insertions and other operations. The best versions of our models outperform previous approaches and set new SOTA on the two Russian GEC benchmarks -- RU-Lang8 and GERA, while achieve competitive performance on RULEC-GEC.

GerMedIQ: At the Gap Between Human and Synthetic Clinical Text

Justin Hofenbitzer, Sebastian Schning, Belle Sebastian, Jacqueline Lammert, Luise Modersohn, Martin Boeker, Diego Frassinelli

Text corpora in non-English clinical contexts are sparse, where synthetic data generation with Large Language Models (LLMs) appears as a promising strategy to overcome this data gap. In order to test the quality of LLMs in generating synthetic data, we applied them to our novel German Medical Interview Questions Corpus (GerMedIQ), consisting of 4,524 unique question-response pairs in German. We augmented our corpus by asking a cohort of models to produce suitable responses to the same questions. Structural and semantic evaluations of the synthetic responses revealed that while augmented responses may meet the grammatical requirements, most models were not able to produce semantically comparable responses to humans. Also, an LLM-as-a-judge experiment showcased that human responses were consistently rated more appropriate than synthetic ones. We find that data augmentation with LLMs in non-English and clinical domain contexts has to be performed carefully.

Mind Flows: A Comparative Study of Narrative Coherence in Human and LLM-Generated Stream-of-Consciousness Essays}

Nellia Dzhubaeva, Katharina Trinley, Laura Pissani

This paper examines differences between stream-of-consciousness narratives written by humans and those generated by large language models (LLMs) to assess narrative coherence and personality expression. We generated texts by prompting LLMs (Llama-3.1-8B \\& DeepSeek-R1-Distill-Llama-8B) with the first half of essays either revealing the personality characteristics (Big Five) or hiding them. Our analysis revealed consistently low similarity between LLM-generated continuations and original human texts, as measured by cosine similarity, perplexity, and BLEU scores. Including explicit personality traits significantly enhanced Llama-3.1-8B's performance, particularly in BLEU scores. Further analysis of personality expression showed varying alignment patterns between LLMs and human texts. Further analysis of personality expression found higher openness for all models. Additionally, Llama-3.1-8B models exhibited higher Extraversion but low Agreeableness, while DeepSeek-R1-Distill-Llama-8B's showed dramatic personality shifts during its thinking process, particularly when provided with personality traits in the prompt.

Exploiting contextual information to improve stance detection in informal political discourse with LLMs

Arman Engin Sucu, Yixiang Zhou

This study investigates the use of Large Language Models (LLMs) for political stance detection in informal online discourse, where language is often sarcastic, ambiguous, and context-dependent. We explore whether providing contextual information, specifically user profile summaries derived from historical posts, can improve classification accuracy. Using a real-world political forum dataset, we generate structured profiles that summarize users' ideological leaning, recurring topics, and linguistic patterns. We evaluate seven state-of-the-art LLMs across baseline and context-enriched setups through a comprehensive cross-model evaluation. Our findings show that contextual prompts significantly boost accuracy, with improvements ranging from +17.5\% to +38.5\%, achieving up to 74\% accuracy that surpasses previous approaches. We also analyze how profile size and post selection strategies affect performance, showing that strategically chosen political content yields better results than larger, randomly selected contexts. These findings underscore the value of incorporating user-level context to enhance LLM performance in nuanced political classification tasks.

Local LLM Zero-Shot Analysis of Homelessness Discourse on Reddit

Jonathan A. Karr Jr., Ben Herbst, Matthew Hauenstein, Georgina Curto, Nitesh V Chawla

Homelessness is a persistent issue, impacting millions worldwide, and over 770,000 people experienced homelessness in the U.S. in 2024. Social stigmatization is a significant barrier to alleviation, shifting public perception, and influencing policy. Online discourse on platforms such as Reddit shape public opinion. To address this, the project leverages natural language processing and large language models (LLMs) to mitigate bias against people experiencing homelessness (PEH) in online spaces. The goal is to promote awareness, reduce harmful biases, inform policy, and improve the fairness of generative AI. We gather Reddit data for 10 U.S. cities, then perform zero-shot classification, and finally, mitigation using Llama 3.2 Instruct and Qwen 2.5 7B Instruct models. The initial results highlighted the differing classifications between models and indicated that many mitigated outputs remained biased. This suggests the need for potential model refinement for the mitigation of text related to PEH.

QA Analysis in Medical and Legal Domains: A Survey of Data Augmentation in Low-Resource Settings

Benedictus Kent Rachmat

Large Language Models (LLMs) have revolutionized Natural Language Processing (NLP), but their success remains largely confined to high-resource, general-purpose domains. In contrast, applying LLMs to low-resource domains poses significant challenges due to limited training data, domain drift, and strict terminology constraints. This survey provides an overview of the current landscape in domain-specific, low-resource QA with LLMs. We begin by analyzing the coverage and representativeness of specialized-domain QA datasets against large-scale reference datasets what we refer to as ParentQA. Building on this analysis, we survey data-centric strategies to enhance input diversity, including data augmentation techniques. We further discuss evaluation metrics for specialized tasks and consider ethical concerns. By mapping current methodologies and outlining open research questions, this survey aims to guide future efforts in adapting LLMs for robust and responsible use in resource-constrained, domain-specific environments.

GenDLN: Evolutionary Algorithm-Based Stacked LLM Framework for Joint Prompt Optimization

Pia Chouayfati, Niklas Herbster, bel Domonkos Sfrn, Matthias Grabmair

With Large Language Model (LLM)-based applications becoming more common due to strong performance across many tasks, prompt optimization has emerged as a way to extract better solutions from frozen, often commercial LLMs that are not specifically adapted to a task. LLM-assisted prompt optimization methods provide a promising alternative to manual/human prompt engineering, where LLM "reasoning" can be used to make them optimizing agents. However, the cost of using LLMs for prompt optimization via commercial APIs remains high, especially for heuristic methods like evolutionary algorithms (EAs), which need many iterations to converge, and thus, tokens, API calls, and rate-limited network overhead. We propose GenDLN, an efficient genetic algorithm-based prompt pair optimization framework that leverages commercial API free tiers. Our approach allows teams with limited resources (NGOs, non-profits, academics. . .) to efficiently use commercial LLMs for EA-based prompt optimization. We conduct experiments on CLAUDETTE for legal terms of service classification and MRPC for paraphrase detection, performing in line with selected prompt optimization baselines, at no cost. Our code is available in <omitted>.

Sign Language Video Segmentation Using Temporal Boundary Identification

Kavu Maithri Rao, HAMIDULLAH Yasser, Eleftherios Avramidis

Sign language segmentation focuses on identifying temporal boundaries within sign language videos. As compared to previous segmentation techniques that have depended on frame-level and phrase-level segmentation, our study emphasizes on subtitle-level segmentation, using synchronized subtitle data to facilitate temporal boundary recognition. Based on Beginning-Inside-Outside (BIO) tagging for subtitle unit delineation, we train a sequence-to-sequence (Seq2Seq) model with and without attention for subtitle boundary identification. Training on optical flow data and aligned subtitles from BOBSL and YouTube-ASL, we show that the Seq2Seq model with attention outperforms baseline models, achieving improved percentage of segments, F1 and IoU score. An additional contribution is the development of an method for subtitle temporal resolution, aiming to facilitate manual annotation.

LIP-NER: Literal Patterns Benefit LLM-Based NER

Ruiqi Li, Li Chen, Zhonghua Yu

Large Language Models (LLMs) can enhance the performance of Named Entity Recognition (NER) tasks by leveraging external knowledge through in-context learning. When it comes to entity-type-related external knowledge, existing methods mainly provide LLMs with semantic information such as the definition and annotation guidelines of an entity type, leaving the effect of orthographic or morphological information on LLM-based NER unexplored. Besides, it is non-trivial to obtain literal patterns written in natural language to serve LLMs. In this work, we propose LiP-NER, an LLM-based NER framework that utilizes \textbf{Li}teral \textbf{P}atterns, the entity-type-related knowledge that directly describes the orthographic and morphological features of entities. We also propose an LLM-based method to automatically acquire literal patterns, which requires only several sample entities rather than any annotation example, thus further reducing human labor. Our extensive experiments suggest that literal patterns can enhance the performance of LLMs in NER tasks. In further analysis, we found that entity types with relatively standardized naming conventions but limited world knowledge in LLMs, as well as entity types with broad and ambiguous names or definitions yet low internal variation among entities, benefit most from our approach. We found that the most effective written literal patterns are (1) detailed in classification, (2) focused on majority cases rather than minorities, and (3) explicit about obvious literal features.

Testing English News Articles for Lexical Homogenization Due to Widespread Use of Large Language Models

Sarah Fitterer, Dominik Gangl, Jannes Ulbrich

It is widely assumed that Large Language Models (LLMs) are shaping language, with multiple studies noting the growing presence of LLM-generated content and suggesting homogenizing effects. However, it remains unclear if these effects are already evident in recent writing. This study addresses that gap by comparing two datasets of English online news articles -- one from 2018, prior to LLM popularization, and one from 2024, after widespread LLM adoption. We define lexical homogenization as a decrease in lexical diversity, measured by the MATTR, Maas, and MTLD metrics, and introduce the LLM-Style-Word Ratio (SWR) to measure LLM influence. We found higher MTLD and SWR scores, yet negligible changes in Maas and MATTR scores in 2024 corpus. We conclude that while there is an apparent influence of LLMs on written online English, homogenization effects do not show in the measurements. We therefore propose to apply different metrics to measure lexical homogenization in future studies on the influence of LLM usage on language change.

Direct Confidence Alignment: Aligning Verbalized Confidence with Internal Confidence In Large Language Models

Glenn Zhang, Treasure Mayowa, Jason Fan, Yicheng Fu, Sean O'Brien, Aaron Sandoval, Kevin Zhu

Producing trustworthy and reliable Large Language Models (LLMs) has become increasingly important as their usage becomes more widespread. Calibration seeks to achieve this by improving the alignment between the model's confidence and the actual likelihood of its responses being correct or desirable. However, it has been observed that the internal confidence of a model, derived from token probabilities, is not well aligned with its verbalized confidence, leading to misleading results with different calibration methods. In this paper, we propose Direct Confidence Alignment (DCA), a method using Direct Preference Optimization to align an LLM's verbalized confidence with its internal confidence rather than ground-truth accuracy, enhancing model transparency and reliability by ensuring closer alignment between the two confidence measures. We evaluate DCA across multiple open-weight LLMs on a wide range of datasets. To further assess this alignment, we also introduce three new calibration error-based metrics. Our results show that DCA improves alignment metrics on certain model architectures, reducing inconsistencies in a model's confidence expression. However, we also show that it can be ineffective on others, highlighting the need for more model-aware approaches in the pursuit of more interpretable and trustworthy LLMs.

Bridging the Data Gap in Financial Sentiment: LLM-Driven Augmentation

Rohit Kumar, Chandan Nolbaria

Static and outdated datasets hinder the accuracy of Financial Sentiment Analysis (FSA) in capturing rapidly evolving market sentiment. We tackle this by proposing a novel data augmentation technique using Retrieval Augmented Generation (RAG). Our method leverages a generative LLM to infuse established benchmarks with up-to-date contextual information from contemporary financial news. While this RAG-based augmentation significantly modernizes the data's alignment with current financial language, we employ a robust BERT-BiGRU judge model to ensure the sentiment of the original annotations is faithfully preserved. Crucially, FSA models trained on this enriched data exhibit enhanced performance on unseen test sets, demonstrating the practical value of our approach for developing more reliable and current sentiment classifiers.

Session 10: Summarization (Tuesday 29 July, 16:00-17:30, Hall 5X)

LCDS: A Logic-Controlled Discharge Summary Generation System Supporting Source Attribution and Expert Review

Cheng Yuan, Xinkai Rui, Yongqi Fan, Yawei Fan, Boyang Zhong, Jiacheng Wang, Weiyan Zhang, Tong Ruan

Despite the remarkable performance of Large Language Models (LLMs) in automated discharge summary generation, they still suffer from generating inaccurate content or fabricating information without valid sources. To address these issues, we propose LCDS, a tool for empowering LLMs with Logic-Controlled Discharge Summary generation. LCDS constructs a source mapping table by calculating the textual similarity between electronic medical records (EMRs) and discharge summaries, providing a structured reference for generation. Based on a comprehensive set of logical rules, LCDS identifies the structured writing logic of discharge summaries and integrates it with EMRs to generate silver discharge summaries. Furthermore, LCDS traces the provenance of generated content, allowing experts to review, provide feedback, and rectify errors to produce golden discharge summaries, which are subsequently recorded for the incremental fine-tuning of LLMs. Our project and demo video are in the GitHub repository https://github.com/ycycyc02/LCDS.

Virtual Presentation

Session 10: V-Presentations (Tuesday 29 July, 16:00-17:30,)

Bel Esprit: Multi-Agent Framework for Building AI Model Pipelines

Yunsu Kim, Ahmedelmogtaba Abdelaziz, Thiago Castro Ferreira, Mohamed Al-Badrashiny, Hassan Sawaf

As the demand for artificial intelligence (AI) grows to address complex real-world tasks, single models are often insufficient, requiring the integration of multiple models into pipelines. This paper introduces Bel Esprit, a conversational agent designed to construct AI model pipelines based on user requirements. Bel Esprit uses a multi-agent framework where subagents collaborate to clarify requirements, build, validate, and populate pipelines with appropriate models. We demonstrate its effectiveness in generating pipelines from ambiguous user queries, using both human-curated and synthetic data. A detailed error analysis highlights ongoing challenges in pipeline building. Bel Esprit is available for a free trial at https://belesprit.aixplain.com.

iPET: An Interactive Emotional Companion Dialogue System with LLM-Powered Virtual Pet World Simulation

Zheyong Xie, Shaosheng Cao, Zuozhu Liu, Zheyu Ye, Zihan Niu, Chonggang Lu, Tong Xu, Enhong Chen, Zhe Xu, Yao Hu, Wei Lu

The rapid advancement of large language models (LLMs) has unlocked transformative potential for role-playing emotional companion products, enabling systems that address emotional well-being, educational goals, and therapeutic applications. However, existing approaches lack sustained personalization or contextual adaptability, limiting their real-world utility. In this paper, we introduce iPET, an LLM-powered pet agent that enhances user engagement by offering rich pet lifestyles and user preference-oriented interactions. The iPet consists of a dialogue module that instantiates virtual pet agents for emotionally interactive conversations, a memory module stores and synthesizes records of agent and user experiences, and a world simulation module generating diverse, preference-driven pet behaviors guided by high-level reflections. Deployed for over 200 days in a real-world non-commercial product, iPET has served millions of users, delivering emotional support to psychologically distressed individuals while validating its effectiveness in practical applications.

MedDecXtract: A Clinician-Support System for Extracting, Visualizing, and Annotating Medical Decisions in Clinical Narratives

Mohamed Elgaar, Hadi Amiri, Mitra Mohtarami, Leo Anthony Celi

Clinical notes contain crucial information about medical decisions, including diagnosis, treatment choices, and follow-up plans. However, these decisions are embedded within unstructured text, making it challenging to systematically analyze decision-making patterns or support clinical workflows. We present MedDecXtract, an open-source interactive system that automatically extracts and visualizes medical decisions from clinical text. The system combines a RoBERTa-based model for identifying ten categories of medical decisions (e.g., diagnosis, treatment, follow-up) according to the DICTUM framework, with an intuitive interface for exploration, visualization, and annotation. The system enables various applications including clinical decision support, research on decision patterns, and creation of training data for improved medical language models. The system and its source code can be accessed at https://mohdelgaar-clinical-decisions.hf.space. A video demo is available at https://youtu.be/19j6-XtIE_s.

RAVEN: Robust Advertisement Video Violation Temporal Grounding via Reinforcement Reasoning

Deyi Ji, Yuekui Yang, Haiyang Wu, Shaoping Ma, Tianrun Chen, Lanyun Zhu

Advertisement (Ad) video violation detection is critical for ensuring platform compliance, but existing methods struggle with precise temporal grounding, noisy annotations, and limited generalization. We propose RAVEN, a novel framework that integrates curriculum reinforcement learning with multimodal large language models (MLLMs) to enhance reasoning and cognitive capabilities for violation detection. RAVEN employs a progressive training strategy, combining precisely and coarsely annotated data, and leverages Group Relative Policy Optimization (GRPO) to develop emergent reasoning abilities without explicit reasoning annotations. Multiple hierarchical sophisticated reward mechanism ensures precise temporal grounding and consistent category prediction. Experiments on industrial datasets and public benchmarks show that RAVEN achieves superior performances in violation category accuracy and temporal interval localization. We also design a pipeline to deploy the RAVEN on the online Ad services, and online A/B testing further validates its practical applicability, with significant improvements in precision and recall. RAVEN also demonstrates strong generalization, mitigating the catastrophic forgetting issue associated with supervised fine-tuning.

DistilQwen2.5: Industrial Practices of Training Distilled Open Lightweight Language Models

Chengyu Wang, Junbing Yan, Yuanhao Yue, Jun Huang

Enhancing computational efficiency and reducing deployment costs for large language models (LLMs) have become critical challenges in various resource-constrained scenarios. In this work, we present DistilQwen2.5, a family of distilled, lightweight LLMs derived from the public Qwen2.5 models. These distilled models exhibit enhanced instruction-following capabilities compared to the original models based on a series of distillation techniques that incorporate knowledge from much larger LLMs. In our industrial practice, we first leverage powerful proprietary LLMs with varying capacities as multi-agent teachers to select, rewrite, and refine instruction-response pairs that are more suitable for student LLMs to learn. After standard fine-tuning, we further leverage a computationally efficient model fusion approach that enables student models to progressively integrate fine-grained hidden knowledge from their teachers. Experimental evaluations demonstrate that the distilled models possess significantly stronger capabilities than their original checkpoints. Additionally, we present use cases to illustrate the applications of our framework in real-world scenarios. To facilitate practical use, we have released all the DistilQwen2.5 models to the open-source community.

SimUSER: Simulating User Behavior with Large Language Models for Recommender System Evaluation

Nicolas Bougie, Narimawa Watanabe

Recommender systems play a central role in numerous real-life applications, yet evaluating their performance remains a significant challenge due to the gap between offline metrics and online behaviors. Given the scarcity and limits (e.g., privacy issues) of real user data, we introduce SimUSER, an agent framework that serves as believable and cost-effective human proxies. SimUSER first identifies self-consistent personas from historical data, enriching user profiles with unique backgrounds and personalities. Then, central to this evaluation are users equipped with persona, memory, perception, and brain modules, engaging in interactions with the recommender system. SimUSER exhibits closer alignment with genuine humans than prior work, both at micro and macro levels. Additionally, we conduct insightful experiments to explore the effects of thumbnails on click rates, the exposure effect, and the impact of reviews on user engagement. Finally, we refine recommender system parameters based on offline A/B test results, resulting in improved user engagement in the real world.

Scaling Context, Not Parameters: Training a Compact 7B Language Model for Efficient Long-Context Processing

Chen Wu, Yin Song

We present MegaBeam-Mistral-7B\footnote{}, a language model that supports 512K-token context length. Our work addresses practical limitations in long-context training, supporting real-world tasks such as compliance monitoring and verification. Evaluated on three long-context benchmarks, our 7B-parameter model demonstrates superior in-context learning performance on HELMET and robust retrieval and tracing capability on RULER. It is currently the only open model to achieve competitive long-range reasoning on BABILong at 512K context length without RAG or targeted fine-tuning. Released as fully open source under the Apache 2.0 license, the model has been downloaded over 100,000 times on Hugging Face.

Towards Multi-System Log Anomaly Detection

Boyang Wang, Runqiang Zang, Hongcheng Guo, Shun Zhang, Shaosheng Cao, Donglin Di, Zhoujun Li

Despite advances in unsupervised log anomaly detection, current models require dataset-specific training, causing costly procedures, limited scalability, and performance bottlenecks. Furthermore, numerous models lack cognitive reasoning abilities, limiting their transferability to similar systems. Additionally, these models often encounter the **"identical shortcut"** predicament, erroneously predicting normal classes when confronted with rare anomaly logs due to reconstruction errors. To address these issues, we propose **MLAD**, a novel **M**ulti-system **L**og **A**nomaly **D**etection model incorporating semantic relational reasoning. Specifically, we extract cross-system semantic patterns and encode them as high-dimensional learnable vectors. Subsequently, we revamp attention formulas to discern keyword significance and model the overall distribution through vector space diffusion. Lastly, we employ a Gaussian mixture model to highlight rare word uncertainty, optimizing the vector space with maximum expectation. Experiments on real-world datasets demonstrate the superiority of MLAD.

ORMind: A Cognitive-Inspired End-to-End Reasoning Framework for Operations Research

Zhiyuan Wang, Bokui Chen, Yinya Huang, Qingxing Cao, Ming He, Jianping Fan, Xiaodan Liang

Operations research (OR) is widely deployed to solve critical decision-making problems with complex objectives and constraints, impacting manufacturing, logistics, finance, and healthcare outcomes. While Large Language Models (LLMs) have shown promising results in various domains, their practical application in industry-relevant operations research (OR) problems presents significant challenges and opportunities. Preliminary industrial applications of LLMs for operations research face two critical deployment challenges: 1) Self-correction focuses on code syntax rather than mathematical accuracy, causing costly errors; 2) Complex expert selection creates unpredictable workflows that reduce transparency and increase maintenance costs, making them impractical for time-sensitive business applications. To address these business limitations, we introduce ORMind, a cognitive-inspired framework that enhances optimization through counterfactual reasoning. Our approach emulates human cognition—implementing an end-to-end workflow that systematically transforms requirements into mathematical models and executable solver code. It is currently being tested internally in Lenovo's AI Assistant, with plans to enhance optimization capabilities for both business and consumer customers. Experiments demonstrate that ORMind outperforms existing methods, achieving a 9.5\% improvement on the NL4Opt dataset and a 14.6\% improvement on the ComplexOR dataset.

RUBRIC-MQM : Span-Level LLM-as-judge in Machine Translation For High-End Models

Ahrii Kim

Referred to as LLM-as-judge, a generative large language model (LLM) has demonstrated considerable efficacy as an evaluator in various tasks, including Machine Translation (MT) by predicting scores or identifying error types for individual sentences. However, its dependability in practical application has yet to be demonstrated, as there is only an approximated match due to the task's open-ended nature. To address this problem, we introduce a straightforward and novel meta-evaluation strategy PromptCUE and evaluate cutting-edge LLM-as-judge models such as GEMBA-MQM. We identify their fundamental deficits, including certain label biases and the inability to assess near-perfect translations. To improve reliability, we investigate more trustworthy and less biased models using multidimensional prompt engineering. Our findings indicate that the combination of span-level error quantification and a rubric-style prompt tailored to the characteristics of LLMs has efficiently addressed the majority of the challenges current LLM-as-judge models face. Furthermore, it demonstrates a considerably enhanced alignment with human values. Accordingly, we present Rubric-MQM, the LLM-as-judge for high-end models and an updated version of GEMBA-MQM.

CoAlign: Uncertainty Calibration of LLM for Geospatial Repartition

Zejun Xie, Zhiqing Hong, Wenjun Lyu, Haotian Wang, Guang Wang, Desheng Zhang

With the rapid expansion of e-commerce and continuous urban evolution, Geospatial Repartition, dividing geographical regions into delivery zones, is essential to optimize various objectives, e.g., on-time delivery rate, for last-mile delivery. Recently, large language models (LLMs) have offered promising capabilities for integrating diverse contextual information that is beneficial for geospatial repartition. However, given the inherent uncertainty in LLMs, adapting them to practical usage in real-world repartition is nontrivial. Thus, we introduce CoAlign, a novel three-stage framework that calibrates LLM uncertainty to enable robust geospatial repartition by transforming the task into a ranking problem, integrating historical data with LLM-generated candidates. It first generates explainable candidate partitions with a multi-criteria strategy and then designs a novel conformal method to rank these candidates relative to historical partitions with coverage guarantees. Finally, CoAlign delivers candidates through an interactive decision support system. Extensive evaluation with real-world data shows that CoAlign effectively calibrates LLM uncertainty and generates partitions that better align with human feedback. Moreover, we have deployed CoAlign in one of the world's largest logistics companies, significantly enhancing their delivery operations by increasing candidate acceptance rates by 300\% and improving on-time delivery rates by 3\%. Our work provides a novel angle to address industrial geospatial decision-making tasks by calibrating LLM uncertainty.

CiteFix: Enhancing RAG Accuracy Through Post-Processing Citation Correction

Harsh Maheshwari, Srikanth Tenneti, Alwarappan Nakkiran

Retrieval Augmented Generation (RAG) has emerged as a powerful application of Large Language Models (LLMs), revolutionizing information search and consumption. RAG systems combine traditional search capabilities with LLMs to generate comprehensive answers to user queries, ideally with accurate citations. However, in our experience of developing a RAG product, LLMs often struggle with source attribution, aligning with other industry studies reporting citation accuracy rates of only about 74\% for popular generative search engines. To address this, we present efficient post-processing algorithms to improve citation accuracy in LLM-generated responses, with minimal impact on latency and cost. Our approaches cross-check generated citations against retrieved articles using methods including keyword + semantic matching, fine tuned model with BERTScore, and a lightweight LLM-based technique. Our experimental results demonstrate a relative improvement of 15.46\% in the overall accuracy metrics of our RAG system. This significant enhancement potentially enables a shift from our current larger language model to a relatively smaller model that is approximately 12x more cost-effective and 3x faster in inference time, while maintaining comparable performance. This research contributes to enhancing the reliability and trustworthiness of AI-generated content in information retrieval and summarization tasks which is critical to gain customer trust especially in commercial products.

Light-R1: Curriculum SFT, DPO and RL for Long COT from Scratch and Beyond

Liang Wen, Yunke Cai, Fenrui Xiao, Xin He, Qi An, zhenyu duan, Yimin Du, Junchen Liu, Tanglifu, Xiaowei Lv, Haosheng Zou, Yongchao DENG, Shousheng Jia, Xiangzheng Zhang

This paper introduces Light-R1, an opensource suite for training long reasoning modelsusing reproducible and cost-effective methodology. Given the proprietary nature of data usedin the DeepSeek-R1 series, we develop an alternative approach leveraging exclusively publicdata and models. Our curriculum training progressively increases data difficulty, combinedwith multi-staged post-training. Our LightR1-32B model, trained from Qwen2.5-32BInstruct, outperforms DeepSeek-R1-DistillQwen-32B in math reasoning. Experimental results show that this curriculum approachbecomes more effective when distinct, diverse datasets are available for different training stages: fine-tuning DeepSeek-R1-Distilledmodels (pre-tuned by DeepSeek team on proprietary data) with 3,000 challenging examplesfrom our curriculum dataset yielded state-ofthe-art 7B and 14B models, while the 32Bmodel, Light-R1-32B-DS performed comparably to QwQ-32B and DeepSeek-R1. Furthermore, we extend our work by applying GRPOon long reasoning models. Our final Light-R1-14B-DS achieves SOTA performance among14B models in math, with AIME24 \& 25 scoresof 74.0 and 60.2 respectively, surpassing many32B models and DeepSeek-R1-Distill-Llama70B. Despite math-focused training, Light-R1-14B-DS demonstrates strong cross-domain generalization. Light-R1 represents a significantadvancement in making sophisticated reasoning models more accessible and implementablein real-world applications. Our models, training data and code have been made available.

Transforming Podcast Preview Generation: From Expert Models to LLM-Based Systems

Winstead Zhu, Ann Clifton, Azin Ghazimatin, Edgar Tanaka, Ward Ronan

Discovering and evaluating long-form talk content such as videos and podcasts poses a significant challenge for users, as it requires a considerable time investment. Previews offer a practical solution by providing concise snippets that showcase key moments of the content, enabling users to make more informed and confident choices. We propose an LLM-based approach for generating podcast episode previews and deploy the solution at scale, serving hundreds of thousands of podcast previews in a real-world application. Comprehensive offline evaluations and online A/B testing demonstrate that LLM-generated previews consistently outperform a strong baseline built on top of various ML expert models, showcasing a significant reduction in the need for meticulous feature engineering. The offline results indicate notable enhancements in understandability, contextual clarity, and interest level, and the online A/B test shows a 4.6% increase in user engagement with preview content, along with a 5x boost in processing efficiency, offering a more streamlined and performant solution compared to the strong baseline of feature-engineered expert models.

EdgeInfinite: A Memory-Efficient Infinite-Context Transformer for Edge Devices

Jiyu Chen, Shuang Peng, Daxiong Luo, Fan Yang, Renshou Wu, Fangyuan Li, Xiaoxin Chen

Transformer-based large language models (LLMs) encounter challenges in processing long sequences on edge devices due to the quadratic complexity of attention mechanisms and growing memory demands from Key-Value (KV) cache. Existing KV cache optimizations struggle with irreversible token eviction in long-output tasks, while alternative sequence modeling architectures prove costly to adopt within established Transformer infrastructure. We present EdgeInfinite, a memory-efficient solution for infinite contexts that integrates compressed memory into Transformer-based LLMs through a trainable memory-gating module. This approach maintains full compatibility with standard Transformer architectures, requiring fine-tuning only a small part of parameters, and enables selective activation of the memory-gating module for long and short context task routing. The experimental result shows that EdgeInfinite achieves comparable performance to baseline Transformer-based LLM on long context benchmarks while optimizing memory consumption and time to first token.

LexGenie: Automated Generation of Structured Reports for European Court of Human Rights Case Law

Santosh T.Y.S.S, Mahmoud Aly, Oana Ichim, Matthias Grabmair

Analyzing large volumes of case law to uncover evolving legal principles, across multiple cases, on a given topic is a demanding task for legal professionals. Structured topical reports provide an effective solution by summarizing key issues, principles, and judgments, enabling comprehensive legal analysis on a particular topic. While prior works have advanced query-based individual case summarization, none have extended to automatically generating multi-case structured reports. To address this, we introduce LexGenie, an automated LLM-based pipeline designed to create structured reports using the entire body of case law on user-specified topics within the European Court of Human Rights jurisdiction. LexGenie retrieves, clusters, and organizes relevant passages by topic to generate a structured outline and cohesive content for each section. Expert evaluation confirms LexGenie's utility in producing structured reports that enhance efficient, scalable legal analysis.

Operational Advice for Dense and Sparse Retrievers: HNSW, Flat, or Inverted Indexes?

Jimmy Lin

Practitioners working on dense retrieval today face a bewildering number of choices. Beyond selecting the embedding model, another consequential choice is the actual implementation of nearest-neighbor vector search. While best practices recommend HNSW indexes, flat vector indexes with brute-force search represent another viable option, particularly for smaller corpora and for rapid prototyping. In this paper, we provide experimental results on the BEIR dataset using the open-source Lucene search library that explicate the tradeoffs between HNSW and flat indexes (including quantized variants) from the perspectives of indexing time, query evaluation performance, and retrieval quality. With additional comparisons between dense and sparse retrievers, our results provide guidance for today's search practitioner in understanding the design space of dense and sparse retrievers. To our knowledge, we are the first to provide operational advice supported by empirical experiments in this regard.

LEAP \& LEAN: Look-ahead Planning and Agile Navigation for LLM Agents

Nikhil Verma, Manasa Bharadwaj

Foundational models endowed with emergent abilities are increasingly deployed as autonomous agents to navigate intricate environments. Despite their capability to comprehend human intentions, even when paired with reasoning traces, they struggle to achieve robust autonomy. In this work, we introduce **LEAP \& LEAN**, a novel paradigm designed to enhance the performance of Large Language Models (LLMs) as autonomous agents. LEAP employs look-ahead planning to refine action selection, while LEAN streamlines navigation through agile prompt construction, enabling more efficient task completion. Together, LEAP \& LEAN address the explore-exploit dilemma, fostering optimal decision-making and improving task performance. We evaluate our framework across diverse, multi-faceted task-oriented domains(WebShop, ALFWorld, and TravelPlanner) using both proprietary and open-source LLM agents. Notably, without any fine-tuning, our framework outperforms agents trained via imitation learning, reinforcement learning, and reasoning-based approaches. Our findings underscore the importance of action and prompt curation to create robust and efficient fully autonomous LLM agents.

MotiR: Motivation-aware Retrieval for Long-Tail Recommendation

Kaichen Zhao, Mingming Li, Haiquan Zhao, Kuien Liu, Zhixu Li, Xueying Li

In the retrieval stage of recommendation systems, two-tower models are widely adopted for their efficiency as a predominant paradigm. However, this method, which relies on collaborative filtering signals, exhibits limitations in modeling similarity for long-tail items. To address this issue, we propose a Motivation-aware Retrieval for Long-Tail Recommendation, named MotiR. The purchase motivations generated by LLMs represent a condensed abstraction of items' intrinsic attributes. By effectively integrating them with traditional item features, this approach enables the two-tower model to capture semantic-level similarities among long-tail items. Furthermore, a gated network-based adaptive weighting mechanism dynamically adjusts representation weights: emphasizing semantic modeling for long-tail items while preserving collaborative signal advantages for popular items. Experimental results demonstrate 60.5% Hit@10 improvements over existing methods on Amazon Books. Industrial deployment in Taobao\&Tmall Group 88VIP scenarios achieves over 4% CTR and CVR improvement, validating the effectiveness of our method.

Enhancing LLM-as-a-Judge through Active-Sampling-based Prompt Optimization

Cheng Zhen, Ervine Zheng, Jilong Kuang, Geoffrey Jay Tso

We introduce an active-sampling-based framework for automatic prompt optimization, designed to enhance the performance of Large Language Model (LLM)-as-a-judge systems, which use LLMs to evaluate the quality of text or other outputs, in label-scarce settings. Unlike existing approaches that rely on extensive annotations, our method starts with no labeled data and iteratively selects and labels a small, diverse, and informative subset of samples to guide prompt refinement. At each iteration, our method evaluates the current prompt based on selected data and automatically updates the prompt, enabling efficient prompt optimization with minimal supervision. Moreover, we formulate sample selection as a convex optimization problem that balances uncertainty and diversity, maximizing the utility of limited labeling budgets. We validate our framework across four popular LLMs and three real-world datasets, including one from a deployed industry product. Results show that our optimized prompts consistently outperform baselines, achieving significant gains in evaluation quality and robustness while substantially reducing labeling costs.

MedPlan: A Two-Stage RAG-Based System for Personalized Medical Plan Generation

Hsin-Ling Hsu, Cong-Tinh Dao, Luning Wang, Zitao Shuai, Thao Nguyen Minh Phan, Jun-En Ding, Chun-Chieh Liao, Pengfei Hu, Xiaoxue Han, Chih-Ho Hsu, Dongsheng Luo, Wen-Chih Peng, Feng Liu, Fang-Ming Hung, Chenwei Wu

Despite recent success in applying large language models (LLMs) to electronic health records (EHR), most systems focus primarily on assessment rather than treatment planning. We identify three critical limitations in current approaches: they generate treatment plans in a single pass rather than following the sequential reasoning process used by clinicians; they rarely incorporate patient-specific historical context; and they fail to effectively distinguish between subjective and objective clinical information. Motivated by the SOAP methodology (Subjective, Objective, Assessment, Plan), we introduce MedPlan, a novel framework that structures LLM reasoning to align with real-life clinician workflows. Our approach employs a two-stage architecture that first generates a clinical assessment based on patient symptoms and objective data, then formulates a structured treatment plan informed by this assessment and enriched with patient-specific information through retrieval-augmented generation. Comprehensive evaluation demonstrates that our method significantly outperforms baseline approaches in both assessment accuracy and treatment plan quality. Our demo system and code are available at https://github.com/JustinHsu1019/MedPlan.

Synthesizing and Adapting Error Correction Data for Mobile Large Language Model Applications

Yanxiang Zhang, Zheng Xu, Shanshan Wu, Yuanbo Zhang, Daniel Ramage

Error correction is an important capability when applying large language models (LLMs) to facilitate user typing on mobile devices. In this paper, we use LLMs to synthesize a high-quality dataset of error correction pairs to evaluate and improve LLMs for mobile applications. We first prompt LLMs with error correction domain knowledge to build a scalable and reliable addition to the existing data synthesis pipeline. We then adapt the synthetic data distribution to match the mobile application domain by reweighting the samples. The reweighting model is learnt by predicting (a handful of) live A/B test metrics when deploying LLMs in production, given the LLM performance on offline evaluation data and scores from a small privacy-preserving on-device language model. Finally, we present best practices for mixing our synthetic data with other data sources to improve model performance on error correction in both offline evaluation and production live A/B testing.

OccuTriage: An AI Agent Orchestration Framework for Occupational Health Triage Prediction

Alok Kumar Sahu, Yi Sun, Eamonn Swanton, Farshid Amirabdollahian, Abi Wren

Occupational Health (OH) triage is a systematic process for evaluating and prioritising workplace health concerns to determine appropriate care and interventions. This research addresses critical triage challenges through our novel AI agent orchestration framework, OccuTriage, developed in collaboration with Healthcare Provider. Our framework simulates healthcare professionals' reasoning using specialized LLM agents, retrieval augmentation with domain-specific knowledge, and a bidirectional decision architecture. Experimental evaluation on 2,589 OH cases demonstrates OccuTriage outperforms single-agent approaches with a 20.16% average discordance rate compared to baseline rates of 43.05%, while matching or exceeding human expert performance (25.11%). The system excels in reducing under-triage rates, achieving 9.84% and 3.1% for appointment and assessor type decisions respectively. These results establish OccuTriage's efficacy in performing complex OH triage while maintaining safety and optimizing resource allocation.

CodeIF: Benchmarking the Instruction-Following Capabilities of Large Language Models for Code Generation

Kaiwen Yan, Hongcheng Guo, Xuanqing Shi, Shaosheng Cao, Donglin Di, Zhoujun Li

With the rapid advancement of Large Language Models (LLMs), the demand for robust instruction-following capabilities in code generation tasks has grown significantly. Code generation not only facilitates faster prototyping and automated testing, but also augments developer efficiency through improved maintainability and reusability of code. In this paper, we introduce CodeIF, the first benchmark specifically designed to assess the abilities of LLMs to adhere to task-oriented instructions within diverse code generation scenarios. CodeIF encompasses a broad range of tasks, including function synthesis, error debugging, algorithmic refactoring, and code explanation, thereby providing a comprehensive suite to evaluate model performance across varying complexity levels and programming domains. We conduct extensive experiments with LLMs, analyzing their strengths and limitations in meeting the demands of these tasks. The experimental results offer valuable insights into how well current models align with human instructions, as well as the extent to which they can generate consistent, maintainable, and contextually relevant code. Our findings not only underscore the critical role that instruction-following LLMs can play in modern software development, but also illuminate pathways for future research aimed at enhancing their adaptability, reliability, and overall effectiveness in automated code generation.

How to Enable Effective Cooperation Between Humans and NLP Models: A Survey of Principles, Formalizations, and Beyond

Chen Huang, Yang Deng, Wenqiang Lei, Jiancheng Lv, Tat-Seng Chua, Jimmy Huang

With the advancement of large language models (LLMs), intelligent models have evolved from mere tools to autonomous agents with their own goals and strategies for cooperating with humans. This evolution has birthed a novel paradigm in NLP, i.e., human-model cooperation, that has yielded remarkable progress in numerous NLP tasks in recent years. In this paper, we take the first step to present a thorough review of human-model cooperation, exploring its principles, formalizations, and open challenges. In particular, we introduce a new taxonomy that provides a unified perspective to summarize existing approaches. Also, we discuss potential frontier areas and their corresponding challenges. We regard our work as an entry point, paving the way for more breakthrough research in this regard.

UniICL: An Efficient ICL Framework Unifying Compression, Selection, and Generation

Jun Gao, Qi Lv, Zili Wang, Tianxiang Wu, Ziqiang Cao, Wenjie Li

In-context learning (ICL) enhances the reasoning abilities of Large Language Models (LLMs) by prepending a few demonstrations. It motivates researchers to introduce more examples to provide additional contextual information for the generation. However, existing methods show a significant limitation due to the problem of excessive growth in context length which causes a large hardware burden. Additionally, shallow-relevant examples selected by out-off-shelf tools hinder LLMs from capturing useful contextual information for generation. In this paper, to approach these limitations, we propose \textbf{UniICL}, a novel Unified ICL framework that unifies demonstration compression, demonstration selection, and final response generation. Furthermore, to avoid repeated compression of the same demonstration and boost inference efficiency, we design a tailored compression strategy that allows UniICL caching compression results into Demonstration Bank(DB). Extensive out-of-domain evaluations prove the advantages of UniICL in both effectiveness and efficiency.

BPP-Search: Enhancing Tree of Thought Reasoning for Mathematical Modeling Problem Solving

Teng Wang, Wing Yin YU, Zhenqi He, Zehua Liu, HaileiGong, Han Wu, Xiongwei Han, Wei Shi, Ruifeng She, Fangzhou Zhu, Tao Zhong

LLMs exhibit advanced reasoning capabilities, offering the potential to transform natural language questions into mathematical models. However, existing open-source datasets in operations research domain lack detailed annotations of the modeling process, such as variable definitions, focusing solely on objective values, which hinders reinforcement learning applications. To address this, we release the StructuredOR dataset, annotated with comprehensive labels that capture the complete mathematical modeling process. We further propose BPP-Search, an algorithm that integrates reinforcement learning into a tree-of-thought structure using Beam search, a Process reward model, and a pairwise Preference algorithm. This approach enables efficient exploration of tree structures, avoiding exhaustive search while improving accuracy. Extensive experiments on StructuredOR, NL4OPT, and MAMO-ComplexLP datasets show that BPP-Search significantly outperforms state-of-the-art methods. In tree-based reasoning, BPP-Search excels in accuracy and efficiency, enabling faster retrieval of correct solutions. The StructuredOR dataset is available on Huggingface https://huggingface.co/datasets/LLM4OR/StructuredOR and GitHub https://github.com/LLM4OR/StructuredOR.

Fusing Highly Specialized Language Models for Comprehensive Expertise

Ning Ding, Yulin Chen, Ganqu Cui, Xingtai Lv, Weilin Zhao, Kaiyan Zhang, Ruobing Xie, Bowen Zhou, Zhiyuan Liu, Maosong Sun

Underlying data distributions of natural language, programming code, and mathematical symbols vary vastly, presenting a complex challenge for large language models (LLMs) that strive to achieve high performance across all three domains simultaneously. Achieving a very high level of proficiency for an LLM within a specific domain often requires extensive training with relevant corpora, which is typically accompanied by a sacrifice in performance in other domains. In this paper, we aim to ``play the dealt cards well'' and propose to fuse models that are already highly-specialized directly. The proposed fusing framework, \modelname, consists of different distinct specialists that are already sufficiently trained on different domains (we mainly focus on language, coding, and mathematics in this paper). A token-level gating mechanism is introduced to blend the specialists' outputs. A two-stage training strategy accompanied by balanced sampling is designed to ensure stability. To effectively train the fused model, we further construct a high-quality supervised instruction tuning dataset, \dataname, which includes text, code, and mathematical content. This dataset comprises approximately 300,000 instructions and covers a wide range of topics in each domain. Experiments show that our model could simultaneously achieve mastery of the three crucial domains.

Re-ranking Using Large Language Models for Mitigating Exposure to Harmful Content on Social Media Platforms

Rajvardhan Oak, Muhammad Haroon, Claire Wonjeong jo, Magdalena Wojcieszak, Anshuman Chhabra

Social media platforms utilize Machine Learning (ML) and Artificial Intelligence (AI) powered recommendation algorithms to maximize user engagement, which can result in inadvertent exposure to harmful content. Current moderation efforts, reliant on classifiers trained with extensive human-annotated data, struggle with scalability and adapting to new forms of harm. To address these challenges, we propose a novel re-ranking approach using Large Language Models (LLMs) in zero-shot and few-shot settings. Our method dynamically assesses and re-ranks content sequences, effectively mitigating harmful content exposure without requiring extensive labeled data. Alongside traditional ranking metrics, we also introduce two new metrics to evaluate the effectiveness of re-ranking in reducing exposure to harmful content. Through experiments on three datasets, three models and across three configurations, we demonstrate that our LLM-based approach significantly outperforms existing proprietary moderation approaches, offering a scalable and adaptable solution for harm mitigation.

Have We Designed Generalizable Structural Knowledge Promptings? Systematic Evaluation and Rethinking

Yichi Zhang, Zhuo Chen, Lingbing Guo, yajing Xu, Shaokai Chen, Mengshu Sun, Binbin Hu, Zhiqiang Zhang, Lei Liang, Wen Zhang, Huajun Chen

Large language models (LLMs) have demonstrated exceptional performance in text generation within current NLP research. However, the lack of factual accuracy is still a dark cloud hanging over the LLM skyscraper. Structural knowledge prompting (SKP) is a prominent paradigm to integrate external knowledge into LLMs by incorporating structural representations, achieving state-of-the-art results in many knowledge-intensive tasks. However, existing methods often focus on specific problems, lacking a comprehensive exploration of the generalization and capability boundaries of SKP. This paper aims to evaluate and rethink the generalization capability of the SKP paradigm from four perspectives including Granularity, Transferability, Scalability, and Universality. To provide a thorough evaluation, we introduce a novel multi-granular, multi-level benchmark called SUBARU, consisting of 9 different tasks with varying levels of granularity and difficulty. Through extensive experiments, we draw key conclusions regarding the generalization of SKP, offering insights to guide the future development and extension of the SKP paradigm.

Inference Compute-Optimal Video Vision Language Models

Peiqi Wang, ShengYun Peng, Xuewen Zhang, Hanchao Yu, Yibo Yang, Lifu Huang, Fujun Liu, Qifan Wang

This work investigates the optimal allocation of inference compute across three key scaling factors in video vision language models: language model size, frame count, and the number of visual tokens per frame. While prior works typically focuses on optimizing model efficiency or improving performance without considering resource constraints, we instead identify optimal model configuration under fixed inference compute budgets. We conduct large-scale training sweeps and careful parametric modeling of task performance to identify the inference compute-optimal frontier. Our experiments reveal how task performance depends on scaling factors and finetuning data size, as well as how changes in data size shift the compute-optimal frontier. These findings translate to practical tips for selecting these scaling factors.

Gender Inclusivity Fairness Index (GIFI): A Multilevel Framework for Evaluating Gender Diversity in Large Language Models

Zhengyang Shan, Emily Diana, Jiawei Zhou

We present a comprehensive evaluation of gender fairness in large language models (LLMs), focusing on their ability to handle both binary and non-binary genders. While previous studies primarily focus on binary gender distinctions, we introduce the Gender Inclusivity Fairness Index (GIFI), a novel and comprehensive metric that quantifies the diverse gender inclusivity of LLMs. GIFI consists of a wide range of evaluations at different levels, from simply probing the model with respect to provided gender pronouns to testing various aspects of model generation and cognitive behaviors under different gender assumptions, revealing biases associated with varying gender identifiers. We conduct extensive evaluations with GIFI on 20 prominent open-source and proprietary LLMs of varying sizes and capabilities, discovering significant variations in LLMs' gender inclusivity. Our study highlights the importance of improving LLMs' inclusivity, providing a critical benchmark for future advancements in gender fairness in generative models.\footnote{Data and code available at .}

MAIN-RAG: Multi-Agent Filtering Retrieval-Augmented Generation

Chia-Yuan Chang, Zhimeng Jiang, Vineeth Rakesh, Menghai Pan, Chin-Chia Michael Yeh, Guanchu Wang, Mingzhi Hu, Zhichao Xu, Yan Zheng, Mahashweta Das, Na Zou

Large Language Models (LLMs) are becoming essential tools for various natural language processing tasks but often suffer from generating outdated or incorrect information. Retrieval-Augmented Generation (RAG) addresses this issue by incorporating external, real-time information retrieval to ground LLM responses. However, the existing RAG systems frequently struggle with the quality of retrieval documents, as irrelevant or noisy documents degrade performance, increase computational overhead, and undermine response reliability. To tackle this problem, we propose Multi-Agent Filtering Retrieval-Augmented Generation (MAIN-RAG), a training-free RAG framework that leverages multiple LLM agents to collaboratively filter and score retrieved documents. Specifically, MAIN-RAG introduces an adaptive filtering mechanism that dynamically adjusts the relevance filtering threshold based on score distributions, effectively minimizing noise while maintaining high recall of relevant documents. The proposed approach leverages inter-agent consensus to ensure robust document selection without requiring additional training data or fine-tuning. Experimental results across four QA benchmarks demonstrate that MAIN-RAG consistently outperforms traditional RAG approaches, achieving a 2–11% improvement in answer accuracy while reducing the number of irrelevant retrieved documents. Quantitative analysis further reveals that our approach achieves superior response consistency and answer accuracy over baseline methods, offering a competitive and practical alternative to training-based solutions.

Towards Better Evaluation for Generated Patent Claims

Lekang Jiang, Pascal A. Scherz, Stefan Goetz

Patent claims define the scope of protection and establish the legal boundaries of an invention. Drafting these claims is a complex and time-consuming process that usually requires the expertise of skilled patent attorneys, which can form a large access barrier for many small enterprises. To solve these challenges, researchers have investigated large language models (LLMs) for automating patent claim generation. However, existing studies highlight inconsistencies between automated evaluation metrics and human expert assessments. To bridge this gap, we introduce Patent-CE, the first comprehensive benchmark for evaluating patent claims. Patent-CE includes comparative claim evaluations annotated by patent experts, focusing on five key criteria: feature completeness, conceptual clarity, terminology consistency, logical linkage, and overall quality. Additionally, we propose PatClaimEval, a novel multi-dimensional evaluation method specifically designed for patent claims. Our experiments demonstrate that PatClaimEval achieves the highest correlation with human expert evaluations across all assessment criteria among all tested metrics. This research provides the groundwork for more accurate evaluations of automated patent claim generation systems.

Enhancing Character-Level Understanding in LLMs through Token Internal Structure Learning

Zhu Xu, Zhiqiang Zhao, Zihan Zhang, Yuchi Liu, Quanwei Shen, Fei Liu, Yu Kuang, Jian He, Conglin Liu

Tokenization methods like Byte-Pair Encoding (BPE) enhance computational efficiency in large language models (LLMs) but often obscure internal character structures within tokens. This limitation hinders LLMs' ability to predict precise character positions, which is crucial in tasks like Chinese Spelling Correction (CSC) where identifying the positions of misspelled characters accelerates correction processes. We propose Token Internal Position Awareness (TIPA), a method that significantly improves models' ability to capture character positions within tokens by training them on reverse character prediction tasks using the tokenizer's vocabulary. Experiments demonstrate that TIPA enhances position prediction accuracy in LLMs, enabling more precise identification of target characters in original text. Furthermore, when applied to downstream tasks that do not require exact position prediction, TIPA still boosts performance in tasks needing character-level information, validating its versatility and effectiveness.

AceEdit: Advancing Continuous Knowledge Editing For Large Language Models

Qi Li, Xiaowen Chu

Knowledge editing (KE) has emerged as a prominent alternative that enables efficient and precise information modification inside language models. However, a critical challenge arises in continuous language models editing — a significant performance decline both in knowledge update and retention when the number of edits increases. By dissecting the perturbation weight of language model in continuous KE, we uncover that disentangled and sparsified knowledge representation can significantly alleviate the performance decline. Building on these insights, we introduce AdaEdit, a novel knowledge editing method. Extensive empirical evaluations on multiple LLMs demonstrate that our proposed methods can enhance the performance of edited LLMs in large-size continuous editing regimes, outperforming existing ones without substantially compromising the general abilities of these models.

Leveraging Human Production-Interpretation Asymmetries to Test LLM Cognitive Plausibility

Suet-Ying Lam, Qingcheng Zeng, Jingyi Wu, Rob Voigt

Whether large language models (LLMs) process language similarly to humans has been the subject of much theoretical and practical debate. We examine this question through the lens of the production-interpretation distinction found in human sentence processing and evaluate the extent to which instruction-tuned LLMs replicate this distinction. Using an empirically documented asymmetry between pronoun production and interpretation in humans for implicit causality verbs as a testbed, we find that some LLMs do quantitatively and qualitatively reflect human-like asymmetries between production and interpretation. We demonstrate that whether this behavior holds depends upon both model size-with larger models more likely to reflect human-like patterns and the choice of meta-linguistic prompts used to elicit the behavior. Our codes and results are available \href{https://github.com/LingMechLab/Production-Interpretation_Asymmetries_ACL2025}{here}.

Improving the Calibration of Confidence Scores in Text Generation Using the Output Distribution's Characteristics

Lorenzo Jaime Yu Flores, Ori Ernst, Jackie CK Cheung

Well-calibrated model confidence scores can improve the usefulness of text generation models. For example, users can be prompted to review predictions with low confidence scores, to prevent models from returning bad or potentially dangerous predictions. However, confidence metrics are not always well calibrated in text generation. One reason is that in generation, there can be many valid answers, which previous methods do not always account for. Hence, a confident model could assign probability to many sequences because they are all valid, and not because it is unsure about how to perform the task. We propose task-agnostic confidence metrics suited to generation, which rely solely on model probabilities without the need for further fine-tuning or heuristics. Using these, we are able to improve the calibration of BART and Flan-T5 on summarization, translation, and question answering datasets.

SafeRAG: Benchmarking Security in Retrieval-Augmented Generation of Large Language Model

Xun Liang, Simin Niu, Zhiyu li, Sensen Zhang, Hanyu Wang, Feiyu Xiong, Zhaoxin Fan, Bo Tang, Jihao Zhao, Jiawei Yang, Shichao Song, Mengwei Wang

The indexing-retrieval-generation paradigm of retrieval-augmented generation (RAG) has been highly successful in solving knowledge-intensive tasks by integrating external knowledge into large language models (LLMs). However, the incorporation of external and unverified knowledge increases the vulnerability of LLMs because attackers can perform attack tasks by manipulating knowledge. In this paper, we introduce a benchmark named SafeRAG designed to evaluate the RAG security. First, we classify attack tasks into silver noise, inter-context conflict, soft ad, and white Denial-of-Service. Next, we construct RAG security evaluation dataset (i.e., SafeRAG dataset) primarily manually for each task. We then utilize the SafeRAG dataset to simulate various attack scenarios that RAG may encounter. Experiments conducted on 14 representative RAG components demonstrate that RAG exhibits significant vulnerability to all attack tasks and even the most apparent attack task can easily bypass existing retrievers, filters, or advanced LLMs, resulting in the degradation of RAG service quality. Code is available at: https://github.com/IAAR-Shanghai/SafeRAG.

Benchmarking Open-ended Audio Dialogue Understanding for Large Audio-Language Models

Kuofeng Gao, Shu-Tao Xia, Ke Xu, Philip Torr, Jindong Gu

Large Audio-Language Models (LALMs), such as GPT-4o, have recently unlocked audio dialogue capabilities, enabling direct spoken exchanges with humans. The potential of LALMs broadens their applicability across a wide range of practical scenarios supported by audio dialogues. However, given these advancements, a comprehensive benchmark to evaluate the performance of LALMs in the open-ended audio dialogue understanding remains absent currently. To address this gap, we propose an **A**udio **D**ialogue **U**nderstanding **Bench**mark **(ADU-Bench),** which consists of 4 benchmark datasets. They assess the open-ended audio dialogue ability for LALMs in 3 general scenarios, 12 skills, 9 multilingual languages, and 4 categories of ambiguity handling. Notably, *we firstly propose the evaluation of ambiguity handling* in audio dialogues that expresses different intentions beyond the same literal meaning of sentences, *e.g.,* `"Really!?"` with different intonations. In summary, ADU-Bench includes over 20,000 open-ended audio dialogues for the assessment of LALMs. Through extensive experiments conducted on 16 LALMs, our analysis reveals that existing LALMs struggle with mathematical symbols and formulas, understanding human behavior such as roleplay, comprehending multiple languages, and handling audio dialogue ambiguities from different phonetic elements, such as intonations, pause positions, and homophones. The benchmark is available at https://adu-bench.github.io/.

Pandora's Box or Aladdin's Lamp: A Comprehensive Analysis Revealing the Role of RAG Noise in Large Language Models

Jinyang Wu, Shuai Zhang, Feihu Che, Mingkuan Feng, Pengpeng Shao, Jianhua Tao

Retrieval-Augmented Generation (RAG) has emerged as a crucial method for addressing hallucinations in large language models (LLMs). While recent research has extended RAG models to complex noisy scenarios, these explorations often confine themselves to limited noise types and presuppose that noise is inherently detrimental to LLMs, potentially deviating from real-world retrieval environments and restricting practical applicability. In this paper, we define seven distinct noise types from a linguistic perspective and establish a Noise RAG Benchmark (NoiserBench), a comprehensive evaluation framework encompassing multiple datasets and reasoning tasks. Through empirical evaluation of eight representative LLMs with diverse architectures and scales, we reveal that these noises can be further categorized into two practical groups: noise that is beneficial to LLMs (aka beneficial noise) and noise that is harmful to LLMs (aka harmful noise). While harmful noise generally impairs performance, beneficial noise may enhance several aspects of model capabilities and overall performance. Our analysis offers insights for developing robust RAG solutions and mitigating hallucinations across diverse retrieval scenarios. Code is available at https://github.com/jinyangwu/NoiserBench.

SECRET: Semi-supervised Clinical Trial Document Similarity Search

Trisha Das, Afrah Shafquat, Mandis Beigi, Jacob Aptekar, Jimeng Sun

Clinical trials are vital for evaluation of safety and efficacy of new treatments. However, clinical trials are resource-intensive, time-consuming and expensive to conduct, where errors in trial design, reduced efficacy, and safety events can result in significant delays, financial losses, and damage to reputation. These risks underline the importance of informed and strategic decisions in trial design to mitigate these risks and improve the chances of a successful trial. Identifying similar historical trials is critical as these trials can provide an important reference for potential pitfalls and challenges including serious adverse events, dosage inaccuracies, recruitment difficulties, patient adherence issues, etc. Addressing these challenges in trial design can lead to development of more effective study protocols with optimized patient safety and trial efficiency. In this paper, we present a novel method to identify similar historical trials by summarizing clinical trial protocols and searching for similar trials based on a query trial's protocol. Our approach significantly outperforms all baselines, achieving up to a 78% improvement in recall@1 and a 53% improvement in precision@1 over the best baseline. We also show that our method outperforms all other baselines in partial trial similarity search and zero-shot patient-trial matching, highlighting its superior utility in these tasks.

Self-Taught Agentic Long Context Understanding

Yufan Zhuang, Xiaodong Yu, Jialian Wu, Ximeng Sun, Ze Wang, Jiang Liu, Yusheng Su, Jingbo Shang, Zicheng Liu, Emad Barsoum

Answering complex, long-context questions remains a major challenge for large language models (LLMs) as it requires effective question clarifications and context retrieval. We propose Agentic Long-Context Understanding (AgenticLU), a framework designed to enhance an LLM's understanding of such queries by integrating targeted self-clarification with contextual grounding within an agentic workflow. At the core of AgenticLU is Chain-of-Clarifications (CoC), where models refine their understanding through self-generated clarification questions and corresponding contextual groundings. By scaling inference as a tree search where each node represents a CoC step, we achieve 97.8% answer recall on NarrativeQA with a search depth of up to three and a branching factor of eight. To amortize the high cost of this search process to training, we leverage the preference pairs for each step obtained by the CoC workflow and perform two-stage model finetuning: (1) supervised finetuning to learn effective decomposition strategies, and (2) direct preference optimization to enhance reasoning quality. This enables AgenticLU models to generate clarifications and retrieve relevant context effectively and efficiently in a single inference pass. Extensive experiments across seven long-context tasks demonstrate that AgenticLU significantly outperforms state-of-the-art prompting methods and specialized long-context LLMs, achieving robust multi-hop reasoning while sustaining consistent performance as context length grows.

Towards Effective and Efficient Continual Pre-training of Large Language Models

Jie Chen, Zhipeng Chen, jiapeng wang, Kun Zhou, Yutao Zhu, Jinhao Jiang, Yingqian Min, Xin Zhao, Zhicheng Dou, Jiaxin Mao, Yankai Lin, Ruihua Song, Jun Xu, Xu Chen, Rui Yan, Zhewei Wei, Di Hu, Wenbing Huang, Ji-Rong Wen

Continual pre-training (CPT) has been an important approach for adapting language models to specific domains or tasks. In this paper, we comprehensively study its key designs to balance the new abilities while retaining the original abilities, and present an effective CPT method that can greatly improve the Chinese language ability and scientific reasoning ability of LLMs. To achieve it, we design specific data mixture and curriculum strategies based on existing datasets and synthetic high-quality data. Concretely, we synthesize multidisciplinary scientific QA pairs based on related web pages to guarantee the data quality, and also devise the performance tracking and data mixture adjustment strategy to ensure the training stability. For the detailed designs, we conduct preliminary studies on a relatively small model, and summarize the findings to help optimize our CPT method. Extensive experiments on a number of evaluation benchmarks show that our approach can largely improve the performance of Llama-3 (8B), including both the general abilities (+8.81 on C-Eval and +6.31 on CMMLU) and the scientific reasoning abilities (+12.00 on MATH and +4.13 on SciEval). Our model, data, and codes are available at https://github.com/RUC-GSAI/Llama-3-SynE.

NGQA: A Nutritional Graph Question Answering Benchmark for Personalized Health-aware Nutritional Reasoning

Zheyuan Zhang, Yiyang Li, Nhi Ha Lan Le, Zehong Wang, Tianyi Ma, Vincent Galassi, Keerthiram Murugesan, Nuno Moniz, Werner Geyer, Nitesh V Chawla, Chuxu Zhang, Yanfang Ye

Diet plays a critical role in human health, yet tailoring dietary reasoning to individual health conditions remains a major challenge. Nutrition Question Answering (QA) has emerged as a popular method for addressing this problem. However, current research faces two critical limitations. On one hand, the absence of datasets involving user-specific medical information severely limits \textit{personalization}. This challenge is further compounded by the wide variability in individual health needs. On the other hand, while large language models (LLMs), a popular solution for this task, demonstrate strong reasoning abilities, they struggle with the domain-specific complexities of personalized healthy dietary reasoning, and existing benchmarks fail to capture these challenges. To address these gaps, we introduce the Nutritional Graph Question Answering (NGQA) benchmark, the first graph question answering dataset designed for personalized nutritional health reasoning. NGQA leverages data from the National Health and Nutrition Examination Survey (NHANES) and the Food and Nutrient Database for Dietary Studies (FNDDS) to evaluate whether a food is healthy for a specific user, supported by explanations of the key contributing nutrients. The benchmark incorporates three question complexity settings and evaluates reasoning across three downstream tasks. Extensive experiments with LLM backbones and baseline models demonstrate that the NGQA benchmark effectively challenges existing models. In sum, NGQA addresses a critical real-world problem while advancing GraphQA research with a novel domain-specific benchmark. Our codebase and dataset are available here.

Understanding Cross-Domain Adaptation in Low-Resource Topic Modeling

Pritom Saha Akash, Kevin Chen-Chuan Chang

Topic modeling plays a vital role in uncovering hidden semantic structures within text corpora, but existing models struggle in low-resource settings where limited target-domain data leads to unstable and incoherent topic inference. We address this challenge by formally introducing domain adaptation for low-resource topic modeling, where a high-resource source domain informs a low-resource target domain without overwhelming it with irrelevant content. We establish a finite-sample generalization bound showing that effective knowledge transfer depends on robust performance in both domains, minimizing latent-space discrepancy, and preventing overfitting to the data. Guided by these insights, we propose DALTA (Domain-Aligned Latent Topic Adaptation), a new framework that employs a shared encoder for domain-invariant features, specialized decoders for domain-specific nuances, and adversarial alignment to selectively transfer relevant information. Experiments on diverse low-resource datasets demonstrate that DALTA consistently outperforms state-of-the-art methods in terms of topic coherence, stability, and transferability.

Uncertainty Propagation on LLM Agent

Qiwei Zhao, Dong Li, Yanchi Liu, Wei Cheng, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Huaxiu Yao, Chen Zhao, Haifeng Chen, Xujiang Zhao

Large language models (LLMs) integrated into multi-step agent systems enable complex decision-making processes across various applications. However, their outputs often lack reliability, making uncertainty estimation crucial. Existing uncertainty estimation methods primarily focus on final-step outputs, which fail to account for cumulative uncertainty over the multi-step decision-making process and the dynamic interactions between agents and their environments. To address these limitations, we propose SAUP (Situation Awareness Uncertainty Propagation), a novel framework that propagates uncertainty through each step of an LLM-based agent's reasoning process. SAUP incorporates situational awareness by assigning situational weights to each step's uncertainty during the propagation. Our method, compatible with various one-step uncertainty estimation techniques, provides a comprehensive and accurate uncertainty measure. Extensive experiments on benchmark datasets demonstrate that SAUP significantly outperforms existing state-of-the-art methods, achieving up to 20% improvement in AUROC.

Efficiently Identifying Watermarked Segments in Mixed-Source Texts

Xuandong Zhao, Chenwen Liao, Yu-Xiang Wang, Lei Li

Text watermarks in large language models (LLMs) are increasingly used to detect synthetic text, mitigating misuse cases like fake news and academic dishonesty. While existing watermarking detection techniques primarily focus on classifying entire documents as watermarked or not, they often neglect the common scenario of identifying individual watermark segments within longer, mixed-source documents. Drawing inspiration from plagiarism detection systems, we propose two novel methods for partial watermark detection. First, we develop a geometry cover detection framework aimed at determining whether there is a watermark segment in long text. Second, we introduce an adaptive online learning algorithm to pinpoint the precise location of watermark segments within the text. Evaluated on three popular watermarking techniques (KGW-Watermark, Unigram-Watermark, and Gumbel-Watermark), our approach achieves high accuracy, significantly outperforming baseline methods. Moreover, our framework is adaptable to other watermarking techniques, offering new insights for precise watermark detection. Our code is publicly available at .

Evaluating Language Models as Synthetic Data Generators

Seungone Kim, Juyoung Suk, Xiang Yue, Vijay Viswanathan, Seongyun Lee, Yizhong Wang, Kiril Gashteovski, Carolin Lawrence, Sean Welleck, Graham Neubig

Given the increasing use of synthetic data in language model (LM) post-training, an LM's ability to generate high-quality data has become nearly as crucial as its ability to solve problems directly. While prior works have focused on developing effective data generation methods, they lack systematic comparison of different LMs as data generators in a unified setting. To address this gap, we propose AgoraBench, a benchmark that provides standardized settings and metrics to evaluate LMs' data generation abilities. Through synthesizing 1.26 million training instances using 6 LMs and training 99 student models, we uncover key insights about LMs' data generation capabilities. First, we observe that LMs exhibit distinct strengths. For instance, GPT-4o excels at generating new problems, while Claude-3.5-Sonnet performs better at enhancing existing ones. Furthermore, our analysis reveals that an LM's data generation ability doesn't necessarily correlate with its problem-solving ability. Instead, multiple intrinsic features of data quality—including response quality, perplexity, and instruction difficulty—collectively serve as better indicators. Finally, we demonstrate that strategic choices in output format and cost-conscious model selection significantly impact data generation effectiveness. Our code, checkpoints, and data are all publicly available at https://github.com/neulab/data-agora.

Learning to Rewrite: Generalized LLM-Generated Text Detection

Ran Li, Wei Hao, Weiliang Zhao, Junfeng Yang, Chengzhi Mao

Detecting text generated by Large Language Models (LLMs) is crucial, yet current detectors often struggle to generalize in open-world settings. We introduce Learning2Rewrite, a novel framework to detect LLM-generated text with exceptional generalization to unseen domains. Capitalized on the finding that LLMs inherently modify LLM-generated content less than human-written text when rewriting, we train an LLM to amplify this disparity, yielding a more distinguishable and generalizable edit distance across diverse text distributions. Extensive experiments on data from 21 independent domains and four major LLMs (GPT-3.5, GPT-4, Gemini, and Llama-3) demonstrate that our detector outperforms state-of-the-art detection methods by up to 23.04\% in AUROC for in-distribution tests, 35.10\% for out-of-distribution tests, and 48.66\% under adversarial attacks. Our unique training objective ensures better generalizability compared to directly training for classification, even when leveraging the same amount of tunable parameters. Our findings suggest that reinforcing LLMs' inherent rewriting tendencies offers a robust and scalable solution for detecting LLM-generated text.

Can Language Models Reason about Individualistic Human Values and Preferences?

Liwei Jiang, Taylor Sorensen, Sydney Levine, Yejin Choi

Recent calls for pluralistic alignment emphasize that AI systems should address the diverse needs of all people. Yet, efforts in this space often require sorting people into fixed buckets of pre-specified diversity-defining dimensions (e.g., demographics), risking smoothing out individualistic variations or even stereotyping. To achieve an authentic representation of diversity that respects individuality, we propose individualistic alignment. While individualistic alignment can take various forms, in this paper, we introduce IndieValueCatalog, a dataset transformed from the influential World Values Survey (WVS), to study language models (LMs) on the specific challenge of individualistic value reasoning. Given a sample of an individual's value-expressing statements, models are tasked with predicting their value judgments in novel cases. With IndieValueCatalog, we reveal critical limitations in frontier LMs' abilities to predict individualistic values with accuracies only ranging between 55% to 65%. Moreover, our results highlight that a precise description of individualistic values cannot be approximated only via demographic information. Finally, we train a series of IndieValueReasoners to reveal new patterns and dynamics into global human values.

SongComposer: A Large Language Model for Lyric and Melody Generation in Song Composition

Shuangrui Ding, Zihan Liu, Xiaoyi Dong, Pan Zhang, Rui Qian, Junhao Huang, Conghui He, Dahua Lin, Jiaqi Wang

Creating lyrics and melodies for the vocal track in a symbolic format, known as song composition, demands expert musical knowledge of melody, an advanced understanding of lyrics, and precise alignment between them. Despite achievements in sub-tasks such as lyric generation, lyric-to-melody, and melody-to-lyric, etc, a unified model for song composition has not yet been achieved. In this paper, we introduce SongComposer, a pioneering step towards a unified song composition model that can readily create symbolic lyrics and melodies following instructions. SongComposer is a music-specialized large language model (LLM) that, for the first time, integrates the capability of simultaneously composing lyrics and melodies into LLMs by leveraging three key innovations: 1) a flexible tuple format for word-level alignment of lyrics and melodies, 2) an extended tokenizer vocabulary for song notes, with scalar initialization based on musical knowledge to capture rhythm, and 3) a multi-stage pipeline that captures musical structure, starting with motif-level melody patterns and progressing to phrase-level structure for improved coherence. Extensive experiments demonstrate that SongComposer outperforms advanced LLMs, including GPT-4, in tasks such as lyric-to-melody generation, melody-to-lyric generation, song continuation, and text-to-song creation. Moreover, we will release SongCompose, a large-scale dataset for training, containing paired lyrics and melodies in Chinese and English.

How to Train Long-Context Language Models (Effectively)

Tianyu Gao, Alexander Wettig, Howard Yen, Danqi Chen

We study continued training and supervised fine-tuning (SFT) of a language model (LM) to make effective use of long-context information. We first establish a reliable evaluation protocol to guide model development—instead of perplexity or simple needle-in-a-haystack (NIAH) tests, we use a broad set of long-context downstream tasks, and we evaluate models after SFT as this better reveals long-context abilities. Supported by our robust evaluations, we run thorough experiments to decide the data mix for continued pre-training, the instruction tuning dataset, and many other design choices such as position extrapolation. We find that (1) code repositories and books are excellent sources of long data, but it is crucial to combine them with high-quality short-context data; (2) training with a sequence length beyond the evaluation length boosts long-context performance; (3) for SFT, using only short instruction datasets yields strong performance on long-context tasks. Our final model, ProLong-8B, which is initialized from Llama-3 and trained on 40B tokens, demonstrates state-of-the-art long-context performance among similarly sized models at a length of 128K. ProLong outperforms Llama-3.1-8B-Instruct on the majority of long-context tasks despite using only 5% as many tokens during long-context training. Additionally, ProLong can effectively process up to 512K tokens, one of the longest context windows of publicly available LMs.

Large Language Models are Good Relational Learners

Fang Wu, Vijay Prakash Dwivedi, Jure Leskovec

Large language models (LLMs) have demonstrated remarkable capabilities across various domains, yet their application to relational deep learning (RDL) remains underexplored. Existing approaches adapt LLMs by traversing relational links between entities in a database and converting the structured data into flat text documents, but this text-based serialization disregards critical relational structures, introduces redundancy, and often exceeds standard LLM context lengths. We introduce Rel-LLM, a novel architecture that employs a graph neural network (GNN) based encoder to create structured relational prompts for LLMs within a retrieval-augmented generation (RAG) framework. Unlike traditional text-based serialization approaches, our method preserves the inherent relational structure of databases while enabling LLMs to effectively process and reason over complex entity relationships. Specifically, the GNN encoder extracts a local subgraph around an entity to build feature representations that contain relevant entity relationships and temporal dependencies. These representations are transformed into structured prompts using a denormalization process, effectively allowing the LLM to reason over relational structures. Through extensive experiments, we demonstrate that Rel-LLM outperforms existing methods on key RDL tasks, offering a scalable and efficient approach to integrating LLMs with structured data sources. Code is available at.

LLM Braces: Straightening Out LLM Predictions with Relevant Sub-Updates

Ying Shen, Lifu Huang

Recent findings reveal that much of the knowledge in a Transformer-based Large Language Model (LLM) is encoded in its feed-forward (FFN) layers, where each FNN layer can be interpreted as the summation of sub-updates, each corresponding to a weighted column vector from the FFN's value parameter matrix that often encodes human-interpretable concepts. In light of this, we hypothesize that model performance and behaviors can be further enhanced and controlled by modulating the contributions of these sub-updates based on their relevance to the input or target output style, and propose LLMBraces, a novel and efficient method that computes relevance scores associated with value vectors in FFN layers and leverages these scores to dynamically adjust the contribution of sub-updates. By optimizing sub-update contributions, LLMBraces refines the prediction process, leading to more accurate and reliable outputs, much like a `brace' providing support and stability. Moreover, LLMBraces can be extended to support conditional control over generation characteristics, such as sentiment, thereby offering fine-grained steering of LLM outputs. Extensive experiments on various LLMs—including Qwen2.5-1.5B, Llama2-7B, and Llama3-8B—demonstrate that LLMBraces outperforms baseline approaches in both fine-tuning and zero-shot settings while requiring significantly fewer tunable parameters, up to 75% fewer compared to LoRA. Furthermore, LLMBraces excels in sentiment-controlled generation and toxicity reduction, highlighting its potential for flexible, controlled text generation across applications.

LLM as Entity Disambiguator for Biomedical Entity-Linking

Christophe Ye, Cassie S. Mitchell

Entity linking involves normalizing a mention in medical text to a unique identifier in a knowledge base, such as UMLS or MeSH. Most entity linkers follow a two-stage process: first, a candidate generation step selects high-quality candidates, and then a named entity disambiguation phase determines the best candidate for final linking. This study demonstrates that leveraging a large language model (LLM) as an entity disambiguator significantly enhances entity linking models' accuracy and recall. Specifically, the LLM disambiguator achieves remarkable improvements when applied to alias-matching entity linking methods. Without any fine-tuning, our approach establishes a new state-of-the-art (SOTA), surpassing previous methods on multiple prevalent biomedical datasets by up to 16 points in accuracy. We released our code on GitHub at https://github.com/ChristopheYe/llm_disambiguator

Exploiting the Shadows: Unveiling Privacy Leaks through Lower-Ranked Tokens in Large Language Models

Yuan Zhou, ZHUO ZHANG, Xiangyu Zhang

Large language models (LLMs) play a crucial role in modern applications but face vulnerabilities related to the extraction of sensitive information. This includes unauthorized accesses to internal prompts and retrieval of personally identifiable information (PII) (e.g., in Retrieval-Augmented Generation based agentic applications). We examine these vulnerabilities in a question-answering (QA) setting where LLMs use retrieved documents or training knowledge as few-shot prompts. Although these documents remain confidential under normal use, adversaries can manipulate input queries to extract private content. In this paper, we propose a novel attack method by exploiting the model's lower-ranked output tokens to leak sensitive information. We systematically evaluate our method, demonstrating its effectiveness in both the agentic application privacy extraction setting and the direct training data extraction. These findings reveal critical privacy risks in LLMs and emphasize the urgent need for enhanced safeguards against information leakage.

Unanswerability Evaluation for Retrieval Augmented Generation

XIANGYU PENG, Prafulla Kumar Choubey, Caiming Xiong, Chien-Sheng Wu

Existing evaluation frameworks for retrieval-augmented generation (RAG) systems focus on answerable queries, but they overlook the importance of appropriately rejecting unanswerable requests. In this paper, we introduce UAEval4RAG, a comprehensive evaluation framework designed to evaluate whether RAG systems effectively handle unanswerable queries specific to a given knowledge base. We first define a taxonomy with six unanswerable categories, and UAEval4RAG automatically synthesizes diverse and challenging queries for any given knowledge base and evaluate the RAG systems with unanswered ratio and acceptable ratio metrics. We also conduct experiments with various RAG components and prompting strategies across four datasets, which reveals that due to varying knowledge distribution across datasets, no single configuration consistently delivers optimal performance on both answerable and unanswerable requests across different knowledge bases. Our findings highlight the critical role of component selection and prompt design in optimizing RAG systems to balance the accuracy of answerable queries with high rejection rates of unanswerable ones. UAEval4RAG provides valuable insights and tools for developing more robust and reliable RAG systems.

SCALE: Towards Collaborative Content Analysis in Social Science with Large Language Model Agents and Human Intervention

Chengshuai Zhao, Zhen Tan, Chau-Wai Wong, Xinyan Zhao, Tianlong Chen, huan liu

Content analysis breaks down complex and unstructured texts into theory-informed numerical categories. Particularly, in social science, this process usually relies on multiple rounds of manual annotation, domain expert discussion, and rule-based refinement. In this paper, we introduce SCALE, a novel multi-agent framework that effectively \underline{\textbf{S}}imulates \underline{\textbf{C}}ontent \underline{\textbf{A}}nalysis via \underline{\textbf{L}}arge language model (LLM) ag\underline{\textbf{E}}nts. SCALE imitates key phases of content analysis, including text coding, collaborative discussion, and dynamic codebook evolution, capturing the reflective depth and adaptive discussions of human researchers. Furthermore, by integrating diverse modes of human intervention, SCALE is augmented with expert input to further enhance its performance. Extensive evaluations on real-world datasets demonstrate that SCALE achieves human-approximated performance across various complex content analysis tasks, offering an innovative potential for future social science research.

SciVer: Evaluating Foundation Models for Multimodal Scientific Claim Verification

Chengye Wang, Yifei Shen, Zexi Kuang, Arman Cohan, Yilun Zhao

We introduce SciVer, the first benchmark specifically designed to evaluate the ability of foundation models to verify claims within a multimodal scientific context. SciVer consists of 3,000 expert-annotated examples over 1,113 scientific papers, covering four subsets, each representing a common reasoning type in multimodal scientific claim verification. To enable fine-grained evaluation, each example includes expert-annotated supporting evidence. We assess the performance of 21 state-of-the-art multimodal foundation models, including o4-mini, Gemini-2.5-Flash, Llama-3.2-Vision, and Qwen2.5-VL. Our experiment reveals a substantial performance gap between these models and human experts on SciVer. Through an in-depth analysis of retrieval-augmented generation (RAG), and human-conducted error evaluations, we identify critical limitations in current open-source models, offering key insights to advance models' comprehension and reasoning in multimodal scientific literature tasks.

MultiAgentBench : Evaluating the Collaboration and Competition of LLM agents

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, Jiaxuan You

Large Language Models (LLMs) have shown remarkable capabilities as autonomous agents; yet existing benchmarks either focus on single-agent tasks or are confined to narrow domains, failing to capture the dynamics of multi-agent coordination and competition. In this paper, we introduce MultiAgentBench, a comprehensive benchmark designed to evaluate LLM-based multi-agent systems across diverse, interactive scenarios. Our framework measures not only task completion but also the quality of collaboration and competition using novel, milestone-based key performance indicators. Moreover, we evaluate various coordination protocols (including star, chain, tree, and graph topologies) and innovative strategies such as group discussion and cognitive planning. Notably, cognitive planning improves milestone achievement rates by 3%. Code and dataset will be made publicly available. Code and datasets are publicavailable at https://github.com/ulab-uiuc/MARBLE

TripleFact: Defending Data Contamination in the Evaluation of LLM-driven Fake News Detection

Cheng Xu, Nan Yan

The proliferation of large language models (LLMs) has introduced unprecedented challenges in fake news detection due to benchmark data contamination (BDC), where evaluation benchmarks are inadvertently memorized during the pre-training, leading to the inflated performance metrics. Traditional evaluation paradigms, reliant on static datasets and closed-world assumptions, fail to account the BDC risk in large-scale pre-training of current LLMs. This paper introduces TripleFact, a novel evaluation framework for fake news detection task, which designed to mitigate BDC risk while prioritizing real-world applicability. TripleFact integrates three components: (1) Human-Adversarial Preference Testing (HAPT) to assess robustness against human-crafted misinformation, (2) Real-Time Web Agent with Asynchronous Validation (RTW-AV) to evaluate temporal generalization using dynamically sourced claims, and (3) Entity-Controlled Virtual Environment (ECVE) to eliminate entity-specific biases. Through experiments on 17 state-of-the-art LLMs, including GPT, LLaMA, and DeepSeek variants, TripleFact demonstrates superior contamination resistance compared to traditional benchmarks. Results reveal that BDC artificially inflates performance by up to 23\% in conventional evaluations, while TripleFact Score (TFS) remain stable within 4\% absolute error under controlled contamination. The framework's ability to disentangle genuine detection capabilities from memorization artifacts underscores its potential as a fake news detection benchmark for the LLM era.

Meaning Beyond Truth Conditions: Evaluating Discourse Level Understanding via Anaphora Accessibility

Xiaomeng Zhu, Zhenghao Zhou, Simon Charlow, Robert Frank

We present a hierarchy of natural language understanding abilities and argue for the importance of moving beyond assessments of understanding at the lexical and sentence levels to the discourse level. We propose the task of anaphora accessibility as a diagnostic for assessing discourse understanding, and to this end, present an evaluation dataset inspired by theoretical research in dynamic semantics. We evaluate human and LLM performance on our dataset and find that LLMs and humans align on some tasks and diverge on others. Such divergence can be explained by LLMs' reliance on specific lexical items during language comprehension, in contrast to human sensitivity to structural abstractions.

Bitnet.cpp: Efficient Edge Inference for Ternary LLMs

Jinheng Wang, Hansong Zhou, Ting Song, Shijie Cao, Yan Xia, Ting Cao, Jianyu Wei, Shuming Ma, Hongyu Wang, Furu Wei

The advent of 1-bit large language models (LLMs), led by BitNet b1.58, has spurred interest in ternary LLMs. Despite this, research and practical applications focusing on efficient edge inference for ternary LLMs remain scarce. To bridge this gap, we introduce Bitnet.cpp, an inference system optimized for BitNet b1.58 and ternary LLMs. Given that mixed-precision matrix multiplication (mpGEMM) constitutes the bulk of inference time in ternary LLMs, Bitnet.cpp incorporates a novel mpGEMM library to facilitate sub-2-bits-per-weight, efficient and lossless inference. The library features two core solutions: Ternary Lookup Table (TL), which addresses spatial inefficiencies of previous bit-wise methods, and Int2 with a Scale (I2_S), which ensures lossless edge inference, both enabling high-speed inference. Our experiments show that Bitnet.cpp achieves up to a 6.25x increase in speed over full-precision baselines and up to 2.32x over low-bit baselines, setting new benchmarks in the field. Additionally, we expand TL to element-wise lookup table (ELUT) for low-bit LLMs in the appendix, presenting both theoretical and empirical evidence of its considerable potential. Bitnet.cpp is publicly available at https://github.com/microsoft/BitNet/tree/paper, offering a sophisticated solution for the efficient and practical deployment of edge LLMs.

KG-Agent: An Efficient Autonomous Agent Framework for Complex Reasoning over Knowledge Graph

Jinhao Jiang, Kun Zhou, Xin Zhao, Yang Song, Chen Zhu, Hengshu Zhu, Ji-Rong Wen

In this paper, we aim to improve the reasoning ability of large language models(LLMs) over knowledge graphs(KGs) to answer complex questions. Inspired by existing methods that design the interaction strategy between LLMs and KG, we propose an autonomous LLM-based agent framework, called \textbf{KG-Agent}, which enables a small LLM to actively make decisions until finishing the reasoning process over KGs. In KG-Agent, we integrate the LLM, multifunctional toolbox, KG-based executor, and knowledge memory, and develop an iteration mechanism that autonomously selects the tool and then updates the memory for reasoning over KG. To guarantee the effectiveness, we leverage program language to formulate the multi-hop reasoning process over the KG and synthesize a code-based instruction dataset to fine-tune the base LLM. Extensive experiments demonstrate that only using 10K samples for tuning LLaMA2-7B can outperform competitive methods using larger LLMs or more data, on both in-domain and out-domain datasets. Our code and data will be publicly released.

Agents Under Siege: Breaking Pragmatic Multi-Agent LLM Systems with Optimized Prompt Attacks

Rana Shahroz, Zhen Tan, Sukwon Yun, Charles Fleming, Tianlong Chen

Most discussions about Large Language Model (LLM) safety have focused on single-agent settings but multi-agent LLM systems now create novel adversarial risks because their behavior depends on communication between agents and decentralized reasoning. In this work, we innovatively focus on attacking pragmatic systems that have constrains such as limited token bandwidth, latency between message delivery, and defense mechanisms. We design a \textit{permutation-invariant adversarial attack} that optimizes prompt distribution across latency and bandwidth-constraint network topologies to bypass distributed safety mechanisms within the system. Formulating the attack path as a problem of \textit{maximum-flow minimum-cost}, coupled with the novel \textit{Permutation-Invariant Evasion Loss (PIEL)}, we leverage {graph-based optimization} to maximize attack success rate while minimizing detection risk. Evaluating across models including \texttt{Llama}, \texttt{Mistral}, \texttt{Gemma}, \texttt{DeepSeek} and other variants on various datasets like \texttt{JailBreakBench} and \texttt{AdversarialBench}, our method outperforms conventional attacks by up to 7\times, exposing critical vulnerabilities in multi-agent systems. Moreover, we demonstrate that existing defenses, including variants of \texttt{Llama-Guard} and \texttt{PromptGuard}, fail to prohibit our attack, emphasizing the urgent need for multi-agent specific safety mechanisms.

A Drop-In Solution for On-the-Fly Adaptation of Speculative Decoding in Large Language Models

Jiesong Liu, Brian Park, Xipeng Shen

Large Language Models (LLMs) are cutting-edge generative AI models built on transformer architecture, which tend to be highly memory-intensive when performing real-time inference. Various strategies have been developed to enhance the end-to-end inference speed for LLMs, one of which is speculative decoding. This technique involves running a smaller LLM (draft model) for inference over a defined window size, denoted as \gamma, while simultaneously being validated by the larger LLM (target model). Choosing the optimal \gamma value and the draft model is essential for unlocking the potential of speculative decoding. But it is difficult to do due to the complicated influence from various factors, including the nature of the task, the hardware in use, and the combination of the large and small models. This paper introduces *on-the-fly adaption of speculative decoding*, a solution that dynamically adapts the choices to maximize the efficiency of speculative decoding for LLM inferences. As a drop-in solution, it needs no offline benchmarking or training. Experiments show that the solution can lead to 3.55-16.48% speed improvement over the standard speculative decoding, and 1.2-3.4\times over the default LLMs.

Think\&Cite: Improving Attributed Text Generation with Self-Guided Tree Search and Progress Reward Modeling

Junyi Li, Hwee Tou Ng

Despite their outstanding capabilities, large language models (LLMs) are prone to hallucination and producing factually incorrect information. This challenge has spurred efforts in attributed text generation, which prompts LLMs to generate content with supporting evidence. In this paper, we propose a novel framework, called Think\&Cite, and formulate attributed text generation as a multi-step reasoning problem integrated with search. Specifically, we propose Self-Guided Monte Carlo Tree Search (SG-MCTS), which capitalizes on the self-reflection capability of LLMs to reason about the intermediate states of MCTS for guiding the tree expansion process. To provide reliable and comprehensive feedback, we introduce Progress Reward Modeling to measure the progress of tree search from the root to the current state from two aspects, i.e., generation and attribution progress. We conduct extensive experiments on three datasets and the results show that our approach significantly outperforms baseline approaches.

Exploring How Generative MLLMs Perceive More Than CLIP with the Same Vision Encoder

Siting Li, Pang Wei Koh, Simon Shaolei Du

Recent research has shown that CLIP models struggle with visual reasoning tasks that require grounding compositionality, understanding spatial relationships, or capturing fine-grained details. One natural hypothesis is that the CLIP vision encoder does not embed essential information for these tasks. However, we find that this is not always the case: The encoder gathers query-relevant visual information, while CLIP fails to extract it. In particular, we show that another branch of Vision-Language Models (VLMs), Generative Multimodal Large Language Models (MLLMs), achieve significantly higher accuracy than CLIP in many of these tasks using the *same* vision encoder and weights, indicating that these Generative MLLMs *perceive more*—as they extract and utilize visual information more effectively. We conduct a series of controlled experiments and reveal that their success is attributed to multiple key design choices, including patch tokens, position embeddings, and prompt-based weighting. On the other hand, enhancing the training data alone or applying a stronger text encoder does not suffice to solve the task, and additional text tokens offer little benefit. Interestingly, we find that fine-grained visual reasoning is not exclusive to generative models trained by an autoregressive loss: When converted into CLIP-like encoders by contrastive finetuning, these MLLMs still outperform CLIP under the same cosine similarity-based evaluation protocol. Our study highlights the importance of VLM architectural choices and suggests directions for improving the performance of CLIP-like contrastive VLMs.

Seeking Rational Demonstrations for Large Language Models: A Domain Generalization Approach to Unsupervised Cross-Domain Keyphrase Generation

Guangzhen Zhao, Yu Yao, Dechang Kong, Zhenjiang Dong

Unsupervised cross-domain keyphrase generation is crucial in real-world natural language processing scenarios. However, the accuracy of up-to-date approaches is limited by the distribution shift between source and target domain, which stems from the cross-domain field. Large language models (LLMs) offer potential for the cross-domain keyphrase generation tasks due to their strong generalization abilities, facilitated by providing demonstrations relevant to the target task. Nevertheless, it is often difficult to obtain labeled samples from the target domain. To address this challenge, this paper aims to seek rational demonstrations from the source domain, thereby improving the LLMs' ability in the unsupervised cross-domain keyphrase generation setting. Specifically, we design a novel domain-aware retrieval model on the source domain. Guided by insights from domain generalization theory, we introduce two generalization terms, one for cross-domain relevance and another for each domain consistency to better support retrieval of rational demonstrations. By the retrieved source-domain demonstrations and distance-based relevant score, the proposed approach achieves optimal accuracy. Comprehensive experiments on widely used cross-domain KG benchmarks demonstrate our approach’s state-of-the-art performance and effectiveness.

LexKeyPlan: Planning with Keyphrases and Retrieval Augmentation for Legal Text Generation: A Case Study on European Court of Human Rights Cases

Santosh T.Y.S.S, Elvin Quero Hernandez

Large language models excel at legal text generation but often produce hallucinations due to their sole reliance on parametric knowledge. Retrieval-augmented models mitigate this by providing relevant external documents to the model but struggle when retrieval is based only on past context, which may not align with the model’s intended future content. We introduce LexKeyPlan, a novel framework that integrates anticipatory planning into generation. Instead of relying solely on context for retrieval, LexKeyPlan generates keyphrases outlining future content serving as forward-looking plan, guiding retrieval for more accurate text generation. This work incorporates planning into legal text generation, demonstrating how keyphrases—representing legal concepts—enhance factual accuracy. By structuring retrieval around legal concepts, LexKeyPlan better aligns with legal reasoning, making it particularly suited for legal applications. Using the ECHR corpus as case study, we show that LexKeyPlan improves factual accuracy and coherence by retrieving information aligned with the intended content.

Unveiling Environmental Impacts of Large Language Model Serving: A Functional Unit View

Yanran Wu, Inez Hua, Yi Ding

Large language models (LLMs) offer powerful capabilities but come with significant environmental impact, particularly in carbon emissions. Existing studies benchmark carbon emissions but lack a standardized basis for comparison across different model configurations. To address this, we introduce the concept of functional unit (FU) as a standardized basis and develop FUEL, the first FU-based framework for evaluating LLM serving's environmental impact. Through three case studies, we uncover key insights and trade-offs in reducing carbon emissions by optimizing model size, quantization strategy, and hardware choice, paving the way for more sustainable LLM serving. The code is available at https://github.com/jojacola/FUEL.

GuessArena: Guess Who I Am? A Self-Adaptive Framework for Evaluating LLMs in Domain-Specific Knowledge and Reasoning

Qingchen Yu, Zifan Zheng, Ding Chen, Simin Niu, Bo Tang, Feiyu Xiong, Zhiyu li

The evaluation of large language models (LLMs) has traditionally relied on static benchmarks, a paradigm that poses two major limitations: (1) predefined test sets lack adaptability to diverse application domains, and (2) standardized evaluation protocols often fail to capture fine-grained assessments of domain-specific knowledge and contextual reasoning abilities. To overcome these challenges, we propose GuessArena, an adaptive evaluation framework grounded in adversarial game-based interactions. Inspired by the interactive structure of the Guess Who I Am? game, our framework seamlessly integrates dynamic domain knowledge modeling with progressive reasoning assessment to improve evaluation fidelity. Empirical studies across five vertical domains-finance, healthcare, manufacturing, information technology, and education-demonstrate that GuessArena effectively distinguishes LLMs in terms of domain knowledge coverage and reasoning chain completeness. Compared to conventional benchmarks, our method provides substantial advantages in interpretability, scalability, and scenario adaptability.

Defining and Evaluating Visual Language Models' Basic Spatial Abilities: A Perspective from Psychometrics

Wenrui Xu, Dalin Lyu, Weihang Wang, Jie Feng, Chen Gao, Yong Li

The Theory of Multiple Intelligences underscores the hierarchical nature of cognitive capabilities. To advance Spatial Artificial Intelligence, we pioneer a psychometric framework defining five Basic Spatial Abilities (BSAs) in Visual Language Models (VLMs): Spatial Perception, Spatial Relation, Spatial Orientation, Mental Rotation, and Spatial Visualization. Benchmarking 13 mainstream VLMs through nine validated psychometric experiments reveals significant gaps versus humans, with three key findings: 1) VLMs mirror human hierarchies (strongest in 2D orientation, weakest in 3D rotation) with independent BSAs; 2) Many smaller models surpass larger counterparts, with Qwen leading and InternVL2 lagging; 3) Interventions like CoT and few-shot training show limits from architectural constraints, while ToT demonstrates the most effective enhancement. Identified barriers include weak geometry encoding and missing dynamic simulation. By linking Psychometrics to VLMs, we provide a comprehensive BSA evaluation benchmark, a methodological perspective for embodied AI development, and a cognitive science-informed roadmap for achieving human-like spatial intelligence.

Enhancing Retrieval Systems with Inference-Time Logical Reasoning

Felix Faltings, Wei Wei, Yujia Bao

Traditional retrieval methods rely on transforming user queries into vector representations and retrieving documents based on cosine similarity within an embedding space. While efficient and scalable, this approach often fails to handle complex queries involving logical constructs such as negations, conjunctions, and disjunctions. In this paper, we propose a novel inference-time logical reasoning framework that explicitly incorporates logical reasoning into the retrieval process. Our method extracts logical reasoning structures from natural language queries and then composes the individual cosine similarity matching scores to formulate the final document scores. This approach enables the retrieval process to handle complex logical reasoning without compromising computational efficiency. Our results on both synthetic and real-world benchmarks demonstrate that the proposed method consistently outperforms traditional retrieval methods across different models and datasets, significantly improving retrieval performance for complex queries.

Using Subtext to Enhance Generative IDRR

Zhipang Wang, Yu Hong, Weihao Sun, Guodong Zhou

Implicit Discourse Relation Recognition (abbr., IDRR) is a NLP task of classifying argument pairs into different types of semantic relations. Arguments contain subtexts, some of which are beneficial to the perception of semantic relations. However, subtexts are connotative. The neural IDRR model fails to be aware of them without being given pertinent prompts. In this paper, we leverage LLaMA to generate subtexts for argument pairs, and verify the effectiveness of subtext-based IDRR. We construct an IDRR baseline using the decoder-only backbone LLaMA, and enhance it with subtext-aware relation reasoning. A confidence-diagnosed dual-channel network is used for collaboration between in-subtext and out-of-subtext IDRR. We experiment on PDTB-2.0 and PDTB-3.0 for both the main-level and secondary-level relation taxonomies. The test results show that our approach yields substantial improvements compared to the baseline, and achieves higher F1-scores on both benchmarks than the previous decoder-only IDRR models. We make the source codes and data publicly available.

Wizard of Shopping: Target-Oriented E-commerce Dialogue Generation with Decision Tree Branching

Xiangci Li, Zhiyu Chen, Jason Ingyu Choi, Nikhita Vedula, Besnik Fetahu, Oleg Rokhlenko, Shervin Malmasi

The goal of conversational product search (CPS) is to develop an intelligent, chat-based shopping assistant that can directly interact with customers to understand shopping intents, ask clarification questions, and find relevant products. However, training such assistants is hindered mainly due to the lack of reliable and large-scale datasets. Prior human-annotated CPS datasets are extremely small in size and lack integration with real-world product search systems. We propose a novel approach, TRACER, which leverages large language models (LLMs) to generate realistic and natural conversations for different shopping domains. TRACER's novelty lies in grounding the generation to dialogue plans, which are product search trajectories predicted from a decision tree model, that guarantees relevant product discovery in the shortest number of search conditions. We also release the first target-oriented CPS dataset Wizard of Shopping (WoS), containing highly natural and coherent conversations (3.6k) from three shopping domains. Finally, we demonstrate the quality and effectiveness of WoS via human evaluations and downstream tasks.

Accelerating Dense LLMs via L0-regularized Mixture-of-Experts

Zhenyu Zhang, JiuDong Yang, taozhaowen, Meng Chen

Large language models (LLMs) achieve strong performance but suffer from slow and costly inference. Existing acceleration methods often lead to noticeable performance degradation, while Mixture-of-Experts (MoE) models require extensive computational resources. In this paper, we propose L0-MoE, a lightweight MoE approach using L0-regularization to accelerate dense LLMs nearly without performance loss. Our method introduces a cluster confusion matrix for domain-aware dataset curation and applies dynamic batching for efficient training. Experiments show that L0-MoE achieves up to 2.5x speedup over dense models while maintaining competitive performance, outperforming existing LLM acceleration baselines.

MadaKV: Adaptive Modality-Perception KV Cache Eviction for Efficient Multimodal Long-Context Inference

Kunxi Li, Zhonghua Jiang, Zhouzhou Shen, ZhaodeWang, chengfei lv, Shengyu Zhang, Fan Wu, Fei Wu

This paper introduces MadaKV, a modality-adaptive key-value (KV) cache eviction strategy designed to enhance the efficiency of multimodal large language models (MLLMs) in long-context inference. In multimodal scenarios, attention heads exhibit varying preferences for different modalities, resulting in significant disparities in modality importance across attention heads. Traditional KV cache eviction methods, which are tailored for unimodal settings, fail to capture modality-specific information, thereby yielding suboptimal performance. MadaKV addresses these challenges through two key components: modality preference adaptation and hierarchical compression compensation. By dynamically sensing modality information within attention heads and adaptively retaining critical tokens, MadaKV achieves substantial reductions in KV cache memory footprint and model inference decoding latency (1.3 to 1.5 times improvement) while maintaining high accuracy across various multimodal long-context tasks. Extensive experiments on representative MLLMs and the MileBench benchmark demonstrate the effectiveness of MadaKV compared to existing KV cache eviction methods.

LexCLiPR: Cross-Lingual Paragraph Retrieval from Legal Judgments

Rohit Upadhya, Santosh T.Y.S.S

Efficient retrieval of pinpointed information from case law is crucial for legal professionals but challenging due to the length and complexity of legal judgments. Existing works mostly often focus on retrieving entire cases rather than precise, paragraph-level information. Moreover, multilingual legal practice necessitates cross-lingual retrieval, most works have been limited to monolingual settings. To address these gaps, we introduce LexCLiPR, a cross-lingual dataset for paragraph-level retrieval from European Court of Human Rights (ECtHR) judgments, leveraging multilingual case law guides and distant supervision to curate our dataset. We evaluate retrieval models in a zero-shot setting, revealing the limitations of pre-trained multilingual models for cross-lingual tasks in low-resource languages and the importance of retrieval based post-training strategies. In fine-tuning settings, we observe that two-tower models excel in cross-lingual retrieval, while siamese architectures are better suited for monolingual tasks. Fine-tuning multilingual models on native language queries improves performance but struggles to generalize to unseen legal concepts, highlighting the need for robust strategies to address topical distribution shifts in the legal queries.

Innovative Image Fraud Detection with Cross-Sample Anomaly Analysis: The Power of LLMs

QiWen Wang, Zhenghao Lin, Chen Lin, Junqi Yang, Zhenzhe Ying, Weiqiang Wang

The financial industry faces a substantial workload in verifying document images. Existing methods based on visual features struggle to identify fraudulent document images due to the lack of visual clues on the tampering region. This paper proposes CSIAD (Cross-Sample Image Anomaly Detection) by leveraging LLMs to identify logical inconsistencies in similar images. This novel framework accurately detects forged images with slight tampering traces and explains anomaly detection results. Furthermore, we introduce CrossCred, a new benchmark of real-world fraudulent images with fine-grained manual annotations. Experiments demonstrate that CSIAD outperforms state-of-the-art image fraud detection methods by 79.6% (F1) on CrossCred and deployed industrial solutions by 21.7% (F1) on business data. The benchmark is available at https://github.com/XMUDM/CSIAD.

RSVP: Reasoning Segmentation via Visual Prompting and Multi-modal Chain-of-Thought

Yi Lu, Jiawang Cao, Yongliang Wu, Bozheng Li, Licheng Tang, Yangguang Ji, Chong Wu, Jay Wu, Wenbo Zhu

Multi-modal Large Language Models (MLLMs) have demonstrated remarkable reasoning capability while lack explicit mechanisms for visual grounding and segmentation, creating a gap between cognitive reasoning and visual perception. To bridge this gap, we introduce Reasoning Segmentation via Visual Prompting (RSVP), a novel framework that unifies multi-step multimodal reasoning with grounded visual understanding. RSVP is a two-stage structuralized framework that integrates reasoning-driven localization with segmentation refinement. In the reasoning stage, RSVP employs multimodal chain-of-thought visual prompts to help MLLMs understand queries and infer targets, generating interpretable region proposals that enhance visual grounding. In segmentation stage, RSVP refines these proposals with a Vision-Language Segmentation Module (VLSM), seamlessly integrates textual and visual cues to produce precise segmentation masks. By explicitly modelling the interaction between multimodal reasoning and segmentation, RSVP introduces a new paradigm for interpretable reasoning segmentation. It exploits MLLMs' inherent localization capabilities, enabling the models to not only reason about objects but also generate structured visual representations. Our extensive experiments demonstrate that RSVP achieves state-of-the-art performance, surpasses state-of-the-art methods by up to +6.5 gIoU and +9.2 cIoU on ReasonSeg, and achieves 49.7 mAP on SegInW under zero-shot settings. These results validate RSVP as an effective and scalable framework for integrating cognitive reasoning with structured visual understanding.

Towards Omni-RAG: Comprehensive Retrieval-Augmented Generation for Large Language Models in Medical Applications

Zhe Chen, Yusheng Liao, Shuyang Jiang, Pingjie Wang, YiQiu Guo, Yanfeng Wang, Yu Wang

Large language models hold promise for addressing medical challenges, such as medical diagnosis reasoning, research knowledge acquisition, clinical decision-making, and consumer health inquiry support. However, they often generate hallucinations due to limited medical knowledge. Incorporating external knowledge is therefore critical, which necessitates multi-source knowledge acquisition. We address this challenge by framing it as a source planning problem, which is to formulate context-appropriate queries tailored to the attributes of diverse sources. Existing approaches either overlook source planning or fail to achieve it effectively due to misalignment between the model's expectation of the sources and their actual content. To bridge this gap, we present MedOmniKB, a repository comprising multigenre and multi-structured medical knowledge sources. Leveraging these sources, we propose the Source Planning Optimisation method, which enhances multi-source utilisation. Our approach involves enabling an expert model to explore and evaluate potential plans while training a smaller model to learn source alignment. Experimental results demonstrate that our method substantially improves multi-source planning performance, enabling the optimised small model to achieve state-of-the-art results in leveraging diverse medical knowledge sources.

Predicting Turn-Taking and Backchannel in Human-Machine Conversations Using Linguistic, Acoustic, and Visual Signals

Yuxin Lin, Yinglin Zheng, Ming Zeng, Wangzheng Shi

This paper addresses the gap in predicting turn-taking and backchannel actions in human-machine conversations using multi-modal signals (linguistic, acoustic, and visual). To overcome the limitation of existing datasets, we propose an automatic data collection pipeline that allows us to collect and annotate over 210 hours of human conversation videos. From this, we construct a Multi-Modal Face-to-Face (MM-F2F) human conversation dataset, including over 1.5M words and corresponding turn-taking and backchannel annotations from approximately 20M frames. Additionally, we present an end-to-end framework that predicts the probability of turn-taking and backchannel actions from multi-modal signals. The proposed model emphasizes the interrelation between modalities and supports any combination of text, audio, and video inputs, making it adaptable to a variety of realistic scenarios. Our experiments show that our approach achieves state-of-the-art performance on turn-taking and backchannel prediction tasks, achieving a 10% increase in F1-score on turn-taking and a 33% increase on backchannel prediction. Our dataset and code are publicly available online to ease of subsequent research.

Rhythm Controllable and Efficient Zero-Shot Voice Conversion via Shortcut Flow Matching

Jialong Zuo, Shengpeng Ji, Minghui Fang, Mingze Li, Ziyue Jiang, Xize Cheng, Xiaoda Yang, Chen Feiyang, Xinyu Duan, Zhou Zhao

Zero-Shot Voice Conversion (VC) aims to transform the source speaker's timbre into an arbitrary unseen one while retaining speech content. Most prior work focuses on preserving the source's prosody, while fine-grained timbre information may leak through prosody, and transferring target prosody to synthesized speech is rarely studied. In light of this, we propose R-VC, a rhythm-controllable and efficient zero-shot voice conversion model. R-VC employs data perturbation techniques and discretize source speech into Hubert content tokens, eliminating much content-irrelevant information. By leveraging a Mask Generative Transformer for in-context duration modeling, our model adapts the linguistic content duration to the desired target speaking style, facilitating the transfer of the target speaker's rhythm. Furthermore, R-VC introduces a powerful Diffusion Transformer (DiT) with shortcut flow matching during training, conditioning the network not only on the current noise level but also on the desired step size, enabling high timbre similarity and quality speech generation in fewer sampling steps, even in just two, thus minimizing latency. Experimental results show that R-VC achieves comparable speaker similarity to state-of-the-art VC methods with a smaller dataset, and surpasses them in terms of speech naturalness, intelligibility and style transfer performance.

Task-Specific Information Decomposition for End-to-End Dense Video Captioning

Zhiyue Liu, Xinru Zhang, Jinyuan Liu

Dense video captioning aims to localize events within input videos and generate concise descriptive texts for each event. Advanced end-to-end methods require both tasks to share the same intermediate features that serve as event queries, thereby enabling the mutual promotion of two tasks. However, relying on shared queries limits the model's ability to extract task-specific information, as event semantic perception and localization demand distinct perspectives on video understanding. To address this, we propose a decomposed dense video captioning framework that derives localization and captioning queries from event queries, enabling task-specific representations while maintaining inter-task collaboration. Considering the roles of different queries, we design a contrastive semantic optimization strategy that guides localization queries to focus on event-level visual features and captioning queries to align with textual semantics. Besides, only localization information is considered in existing methods for label assignment, failing to ensure the relevance of the selected queries to descriptions. We jointly consider localization and captioning losses to achieve a semantically balanced assignment process. Extensive experiments on the YouCook2 and ActivityNet Captions datasets demonstrate that our framework achieves state-of-the-art performance.

Scaling Text-Rich Image Understanding via Code-Guided Synthetic Multimodal Data Generation

Yue Yang, Ajay Patel, Matt Deitke, Tanmay Gupta, Luca Weihs, Andrew Head, Mark Yatskar, Chris Callison-Burch, Ranjay Krishna, Aniruddha Kembhavi, Christopher Clark

Reasoning about images with rich text, such as charts and documents, is a critical application of vision-language models (VLMs). However, VLMs often struggle in these domains due to the scarcity of diverse text-rich vision-language data. To address this challenge, we present CoSyn, a framework that leverages the coding capabilities of text-only large language models (LLMs) to automatically create synthetic text-rich multimodal data. Given input text describing a target domain (e.g., "nutrition fact labels"), CoSyn prompts an LLM to generate code (Python, HTML, LaTeX, etc.) for rendering synthetic images. With the underlying code as textual representations of the synthetic images, CoSyn can generate high-quality instruction-tuning data, again relying on a text-only LLM. Using CoSyn, we constructed a dataset comprising 400K images and 2.7M rows of vision-language instruction-tuning data. Comprehensive experiments on seven benchmarks demonstrate that models trained on our synthetic data achieve state-of-the-art performance among competitive open-source models, including Llama 3.2, and surpass proprietary models such as GPT-4V and Gemini 1.5 Flash. Furthermore, CoSyn can produce synthetic pointing data, enabling VLMs to ground information within input images, showcasing its potential for developing multimodal agents capable of acting in real-world environments.

ToolCoder: A Systematic Code-Empowered Tool Learning Framework for Large Language Models

Hanxing Ding, Shuchang Tao, Liang Pang, Zihao Wei, Jinyang Gao, Bolin Ding, Huawei Shen, Xueqi Cheng

Tool learning has emerged as a crucial capability for large language models (LLMs) to solve complex real-world tasks through interaction with external tools. Existing approaches face significant challenges, including reliance on hand-crafted prompts, difficulty in multi-step planning, and lack of precise error diagnosis and reflection mechanisms. We propose \texttt{ToolCoder}, a novel framework that reformulates tool learning as a code generation task. Inspired by software engineering principles, \texttt{ToolCoder} transforms natural language queries into structured Python function scaffold and systematically breaks down tasks with descriptive comments, enabling LLMs to leverage coding paradigms for complex reasoning and planning. It then generates and executes function implementations to obtain final responses. Additionally, \texttt{ToolCoder} stores successfully executed functions in a repository to promote code reuse, while leveraging error traceback mechanisms for systematic debugging, optimizing both execution efficiency and robustness. Experiments demonstrate that \texttt{ToolCoder} achieves superior performance in task completion accuracy and execution reliability compared to existing approaches, establishing the effectiveness of code-centric approaches in tool learning.

MemeQA: Holistic Evaluation of Meme Understanding

Khoi P. N. Nguyen, Terrence Li, Derek Lou Zhou, Gabriel Xiong, Pranav Balu, Nandhan Alahari, Alan Huang, Tanush Chauhan, Harshavardhan Bala, Emre Guzelordu, Affan Kashfi, Aaron Xu, Suyesh Shrestha, Megan Vu, Jerry Wang, Vincent Ng

Automated meme understanding requires systems to demonstrate fine-grained visual recognition, commonsense reasoning, and extensive cultural knowledge. However, existing benchmarks for meme understanding only concern narrow aspects of meme semantics. To fill this gap, we present MemeQA, a dataset of over 9,000 multiple-choice questions designed to holistically evaluate meme comprehension across seven cognitive aspects. Experiments show that state-of-the-art Large Multimodal Models perform much worse than humans on MemeQA. While fine-tuning improves their performance, they still make many errors on memes wherein proper understanding requires going beyond surface-level sentiment. Moreover, injecting "None of the above" into the available options makes the questions more challenging for the models. Our dataset is publicly available at https://github.com/npnkhoi/memeqa.

CoCoLex: Confidence-guided Copy-based Decoding for Grounded Legal Text Generation

Santosh T.Y.S.S, Youssef Tarek Elkhayat, Oana Ichim, Pranav Shetty, Dongsheng Wang, Zhiqiang Ma, Armineh Nourbakhsh, Xiaomo Liu

Due to their ability to process long and complex contexts, LLMs can offer key benefits to the Legal domain, but their adoption has been hindered by their tendency to generate unfaithful, ungrounded, or hallucinatory outputs. While Retrieval-Augmented Generation offers a promising solution by grounding generations in external knowledge, it offers no guarantee that the provided context will be effectively integrated. To address this, context-aware decoding strategies have been proposed to amplify the influence of relevant context, but they usually do not explicitly enforce faithfulness to the context. In this work, we introduce Confidence-guided Copy-based Decoding for Legal Text Generation (CoCoLex)—a decoding strategy that dynamically interpolates the model produced vocabulary distribution with a distribution derived based on copying from the context. CoCoLex encourages direct copying based on models' confidence, ensuring greater fidelity to the source. Experimental results on five legal benchmarks demonstrate that CoCoLex outperforms existing context-aware decoding methods, particularly in long-form generation tasks.

Vulnerability of LLMs to Vertically Aligned Text Manipulations

Zhecheng Li, Yiwei Wang, Bryan Hooi, Yujun Cai, Zhen Xiong, Nanyun Peng, Kai-Wei Chang

Vertical text input is commonly encountered in various real-world applications, such as mathematical computations and word-based Sudoku puzzles. While current large language models (LLMs) have excelled in natural language tasks, they remain vulnerable to variations in text formatting.Recent research demonstrates that modifying input formats, such as vertically aligning words for encoder-based models, can substantially lower accuracy in text classification tasks. While easily understood by humans, these inputs can significantly mislead models, posing a potential risk of bypassing detection in real-world scenarios involving harmful or sensitive information. With the expanding application of LLMs, a crucial question arises: Do decoder-based LLMs exhibit similar vulnerabilities to vertically formatted text input? In this paper, we investigate the impact of vertical text input on the performance of various LLMs across multiple text classification datasets and analyze the underlying causes. Our findings are as follows: (i) Vertical text input significantly degrades the accuracy of LLMs in text classification tasks. (ii) Chain of Thought (CoT) reasoning does not help LLMs recognize vertical input or mitigate its vulnerability, but \textit{few-shot learning} with careful analysis does. (iii) We explore the underlying cause of the vulnerability by analyzing the inherent issues in tokenization and attention matrices.

Generalized Attention Flow: Feature Attribution for Transformer Models via Maximum Flow

Behrooz Azarkhalili, Maxwell W. Libbrecht

This paper introduces Generalized Attention Flow (GAF), a novel feature attribution method for Transformer-based models to address the limitations of current approaches. By extending Attention Flow and replacing attention weights with the generalized Information Tensor, GAF integrates attention weights, their gradients, the maximum flow problem, and the barrier method to enhance the performance of feature attributions. The proposed method exhibits key theoretical properties and mitigates the shortcomings of prior techniques that rely solely on simple aggregation of attention weights. Our comprehensive benchmarking on sequence classification tasks demonstrates that a specific variant of GAF consistently outperforms state-of-the-art feature attribution methods in most evaluation settings, providing a more reliable interpretation of Transformer model outputs.

Improving Language and Modality Transfer in Translation by Character-level Modeling

Ioannis Tsiamas, David Dale, Marta R. Costa-jussà

Current translation systems, despite being highly multilingual, cover only 5\% of the world's languages. Expanding language coverage to the long-tail of low-resource languages requires data-efficient methods that rely on cross-lingual and cross-modal knowledge transfer. To this end, we propose a character-based approach to improve adaptability to new languages and modalities. Our method leverages SONAR, a multilingual fixed-size embedding space with different modules for encoding and decoding. We use a teacher-student approach with parallel translation data to obtain a character-level encoder. Then, using ASR data, we train a lightweight adapter to connect a massively multilingual CTC ASR model (MMS), to the character-level encoder, potentially enabling speech translation from 1,000+ languages. Experimental results in text translation for 75 languages on FLORES+ demonstrate that our character-based approach can achieve better language transfer than traditional subword-based models, especially outperforming them in low-resource settings, and demonstrating better zero-shot generalizability to unseen languages. Our speech adaptation, maximizing knowledge transfer from the text modality, achieves state-of-the-art results in speech-to-text translation on the FLEURS benchmark on 33 languages, surpassing previous supervised and cascade models, albeit being a zero-shot model with minimal supervision from ASR data.

Error-driven Data-efficient Large Multimodal Model Tuning

Barry Menglong Yao, Qifan Wang, Lifu Huang

Large Multimodal Models (LMMs) have demonstrated impressive performance across numerous academic benchmarks. However, fine-tuning still remains essential to achieve satisfactory performance on downstream tasks, while the task-specific tuning samples are usually not readily available or expensive and time-consuming to obtain. To address this, we propose an error-driven data-efficient tuning framework that aims to efficiently adapt generic LMMs to newly emerging tasks without requiring extensive task-specific training samples. In our approach, a generic LMM, acting as a student model, is first evaluated on a small validation set of the target task, and then a more powerful model, acting as a teacher model, identifies the erroneous steps within the student model's reasoning steps and analyzes its capability gaps from fully addressing the target task. Based on these gaps, targeted training samples are further retrieved from existing task-agnostic datasets to tune the student model and tailor it to the target task. We perform extensive experiments across three different training data scales and seven tasks, demonstrating that our training paradigm significantly and efficiently improves LMM's performance on downstream tasks, achieving an average performance boost of 7.01%

Planning with Diffusion Models for Target-Oriented Dialogue Systems

Hanwen Du, Bo Peng, Xia Ning

Target-Oriented Dialogue (TOD) remains a significant challenge in the LLM era, where strategic dialogue planning is crucial for directing conversations toward specific targets. However, existing dialogue planning methods generate dialogue plans in a step-by-step sequential manner, and may suffer from compounding errors and myopic actions. To address these limitations, we introduce a novel dialogue planning framework, DiffTOD, which leverages diffusion models to enable non-sequential dialogue planning. DiffTOD formulates dialogue planning as a trajectory generation problem with conditional guidance, and leverages a diffusion language model to estimate the likelihood of the dialogue trajectory. To optimize the dialogue action strategies, DiffTOD introduces three tailored guidance mechanisms for different target types, offering flexible guidance toward diverse TOD targets at test time. Extensive experiments across three diverse TOD settings show that DiffTOD can effectively perform non-myopic lookahead exploration and optimize action strategies over a long horizon through non-sequential dialogue planning, and demonstrates strong flexibility across complex and diverse dialogue scenarios. Our code and data are accessible through https://github.com/ninglab/DiffTOD.

Interactive and Expressive Code-Augmented Planning with Large Language Models

Anthony Zhe Liu, Xinhe Wang, Jacob Sansom, Yao Fu, Jongwook Choi, Sungryull Sohn, Jaekyeom Kim, Honglak Lee

Large Language Models (LLMs) demonstrate strong abilities in common-sense reasoning and interactive decision-making, but often struggle with complex, long-horizon planning tasks. Recent techniques have sought to structure LLM outputs using control flow and code to improve planning performance. However, code-based approaches can be error-prone and insufficient for handling ambiguous or unstructured data. To address these challenges, we propose REPL-Plan, an LLM planning approach that is fully code-expressive (it can utilize all the benefits of code) while also being dynamic (it can flexibly adapt from errors and use the LLM for soft reasoning). In REPL-Plan, an LLM solves tasks by interacting with a Read-Eval-Print Loop (REPL), which iteratively executes and evaluates code, similar to language shells or interactive code notebooks, allowing the model to flexibly correct errors and handle tasks dynamically. We demonstrate that REPL-Plan achieves strong results across various planning domains compared to previous methods.

Enhancing Human Evaluation in Machine Translation with Comparative Judgement

Yixiao Song, Parker Riley, Daniel Deutsch, Markus Freitag

Human evaluation is crucial for assessing rapidly evolving language models but is influenced by annotator proficiency and task design. This study explores the integration of comparative judgment into human annotation for machine translation (MT) and evaluates three annotation setups—point-wise Multidimensional Quality Metrics (MQM), side-by-side (S×S) MQM, and its simplified version S×S relative ranking (RR). In MQM, annotators mark error spans with categories and severity levels. S×S MQM extends MQM to pairwise error annotation for two translations of the same input, while S×S RR focuses on selecting the better output without labeling errors. Key findings are: (1) the S×S settings achieve higher inter-annotator agreement than MQM; (2) S×S MQM enhances inter-translation error marking consistency compared to MQM by, on average, 38.5% for explicitly compared MT systems and 19.5% for others; (3) all annotation settings return stable system rankings, with S×S RR offering a more efficient alternative to (S×S) MQM; (4) the S×S settings highlight subtle errors overlooked in MQM without altering absolute system evaluations. To spur further research, we will release the triply annotated datasets comprising 377 ZhEn and 104 EnDe annotation examples, each covering 10 systems.

Improved Unbiased Watermark for Large Language Models

Ruibo Chen, Yihan Wu, Junfeng Guo, Heng Huang

As artificial intelligence surpasses human capabilities in text generation, the necessity to authenticate the origins of AI-generated content has become paramount. Unbiased watermarks offer a powerful solution by embedding statistical signals into language model-generated text without distorting the quality. In this paper, we introduce MCmark, a family of unbiased, Multi-Channel-based watermarks. MCmark works by partitioning the model's vocabulary into segments and promoting token probabilities within a selected segment based on a watermark key. We demonstrate that MCmark not only preserves the original distribution of the language model but also offers significant improvements in detectability and robustness over existing unbiased watermarks. Our experiments with widely-used language models demonstrate an improvement in detectability of over 10% using MCmark, compared to existing state-of-the-art unbiased watermarks. This advancement underscores MCmark's potential in enhancing the practical application of watermarking in AI-generated texts.

VQ-Eval: Evaluating Multimodal LLMs for Generating Feedback on AIGC Videos

Tingyu Song, Guo Gan, Tongyan Hu, Yilun Zhao

Recently, multimodal large language models (MLLMs) have been extensively explored in video question answering. However, most existing assessments focus on natural videos, overlooking synthetic videos (e.g., AI-generated content). Meanwhile, some works in video generation rely on MLLMs to evaluate the quality of generated videos, but the capabilities of MLLMs on AIGC videos remain largely underexplored. To address this, we propose a new benchmark, VQ-Eval, which introduces four tasks—coherence validation, error awareness, error type detection, and reasoning evaluation—to comprehensively evaluate the abilities of MLLMs on AIGC videos. We evaluate 13 frontier MLLMs on VQ-Eval and find that even the best-performing model, GPT-4.1, struggles to achieve consistently good performance across all tasks. This highlights the challenging nature of our benchmark. Additionally, to investigate the practical applications of VQ-Eval in improving video generation, we design a re-prompt pipeline, demonstrating that aligning MLLMs more closely with human feedback can benefit the video generation.

Typology-Guided Adaptation for African NLP

Ndapa Nakashole

Multilingual models often treat language diversity as a problem of data imbalance, overlooking structural variation. We introduce the *Morphological Index* (MoI), a typologically grounded metric that quantifies how strongly a language relies on surface morphology for noun classification. Building on MoI, we propose *MoI-MoE*, a Mixture of Experts model that routes inputs based on morphological structure. Evaluated on 10 Bantu languages—a large, morphologically rich and underrepresented family—MoI-MoE outperforms strong baselines, improving Swahili accuracy by 14 points on noun class recognition while maintaining performance on morphology-rich languages like Zulu. These findings highlight typological structure as a practical and interpretable signal for multilingual model adaptation.

Evaluating Visual and Cultural Interpretation: The K-Viscuit Benchmark with Human-VLM Collaboration

ChaeHun Park, Yujin Baek, Jaeseok Kim, Yu-Jung Heo, Du-Seong Chang, Jaegul Choo

To create culturally inclusive vision-language models (VLMs), developing a benchmark that tests their ability to address culturally relevant questions is essential. Existing approaches typically rely on human annotators, making the process labor-intensive and creating a cognitive burden in generating diverse questions. To address this, we propose a semi-automated framework for constructing cultural VLM benchmarks, specifically targeting multiple-choice QA. This framework combines human-VLM collaboration, where VLMs generate questions based on guidelines, a small set of annotated examples, and relevant knowledge, followed by a verification process by native speakers. We demonstrate the effectiveness of this framework through the creation of K-Viscuit, a dataset focused on Korean culture. Our experiments on this dataset reveal that open-source models lag behind proprietary ones in understanding Korean culture, highlighting key areas for improvement. We also present a series of further analyses, including human evaluation, augmenting VLMs with external knowledge, and the evaluation beyond multiple-choice QA. Our dataset is available at https://huggingface.co/datasets/ddehun/k-viscuit.

ProMALex: Progressive Modular Adapters for Multi-Jurisdictional Legal Language Modeling

Santosh T.Y.S.S, Mohamed Hesham Elganayni

This paper addresses the challenge of adapting language models to the jurisdiction-specific nature of legal corpora. Existing approaches—training separate models for each jurisdiction or using a single shared model—either fail to leverage common legal principles beneficial for low-resource settings or risk negative interference from conflicting jurisdictional interpretations. To overcome these limitations, we propose a parameter-efficient framework ProMALex, that first derives hierarchical relationships across jurisdictions and progressively inserts adapter modules across model layers based on jurisdictional similarity. This design allows modules in lower layers to be shared across jurisdictions, capturing common legal principles, while higher layers specialize through jurisdiction-specific adapters. Experimental results on two legal language modeling benchmarks demonstrate that ProMALex outperforms both fully shared and jurisdiction-specific models.

Don't Get Lost in the Trees: Streamlining LLM Reasoning by Overcoming Tree Search Exploration Pitfalls

Ante Wang, Linfeng Song, Ye Tian, Dian Yu, Haitao Mi, Xiangyu Duan, Zhaopeng Tu, Jinsong Su, Dong Yu

Recent advancements in tree search algorithms guided by verifiers have significantly enhanced the reasoning capabilities of large language models (LLMs), but at the cost of increased computational resources. In this work, we identify two key challenges contributing to this inefficiency: \textit{over-exploration} due to redundant states with semantically equivalent content, and \textit{under-exploration} caused by high variance in verifier scoring leading to frequent trajectory switching. To address these issues, we propose FETCH -- an e{\bf f}fici{\bf e}nt {\bf t}ree sear{\bf ch} framework, which is a flexible, plug-and-play system compatible with various tree search algorithms. Our framework mitigates over-exploration by merging semantically similar states using agglomerative clustering of text embeddings obtained from a fine-tuned SimCSE model. To tackle under-exploration, we enhance verifiers by incorporating temporal difference learning with adjusted \lambda-returns during training to reduce variance, and employing a verifier ensemble to aggregate scores during inference. Experiments on GSM8K, GSM-Plus, and MATH datasets demonstrate that our methods significantly improve reasoning accuracy and computational efficiency across four different tree search algorithms, paving the way for more practical applications of LLM-based reasoning. The code is available at https://github.com/DeepLearnXMU/Fetch.

Towards Dynamic Theory of Mind: Evaluating LLM Adaptation to Temporal Evolution of Human States

Yang Xiao, Jiashuo WANG, Qiancheng Xu, Changhe Song, Chunpu Xu, Yi Cheng, Wenjie Li, Pengfei Liu

As Large Language Models (LLMs) increasingly participate in human-AI interactions, evaluating their Theory of Mind (ToM) capabilities - particularly their ability to track dynamic mental states - becomes crucial. While existing benchmarks assess basic ToM abilities, they predominantly focus on static snapshots of mental states, overlooking the temporal evolution that characterizes real-world social interactions. We present **DynToM**, a novel benchmark specifically designed to evaluate LLMs' ability to understand and track the temporal progression of mental states across interconnected scenarios. Through a systematic four-step framework, we generate 1,100 social contexts encompassing 5,500 scenarios and 78,100 questions, each validated for realism and quality. Our comprehensive evaluation of ten state-of-the-art LLMs reveals that their average performance underperforms humans by 44.7%, with performance degrading significantly when tracking and reasoning about the shift of mental states. This performance gap highlights fundamental limitations in current LLMs' ability to model the dynamic nature of human mental states.

InstructPart: Task-Oriented Part Segmentation with Instruction Reasoning

Zifu Wan, Yaqi Xie, Ce Zhang, Zhiqiu Lin, Zihan Wang, Simon Stepputtis, Deva Ramanan, Katia P. Sycara

Large multimodal foundation models, particularly in the domains of language and vision, have significantly advanced various tasks, including robotics, autonomous driving, information retrieval, and grounding. However, many of these models perceive objects as indivisible, overlooking the components that constitute them. Understanding these components and their associated affordances provides valuable insights into an object's functionality, which is fundamental for performing a wide range of tasks. In this work, we introduce a novel real-world benchmark, InstructPart, comprising hand-labeled part segmentation annotations and task-oriented instructions to evaluate the performance of current models in understanding and executing part-level tasks within everyday contexts. Through our experiments, we demonstrate that task-oriented part segmentation remains a challenging problem, even for state-of-the-art Vision-Language Models (VLMs). In addition to our benchmark, we introduce a simple baseline that achieves a twofold performance improvement through fine-tuning with our dataset. With our dataset and benchmark, we aim to facilitate research on task-oriented part segmentation and enhance the applicability of VLMs across various domains, including robotics, virtual reality, information retrieval, and other related fields. Project website: https://zifuwan.github.io/InstructPart/.

STaR-SQL: Self-Taught Reasoner for Text-to-SQL

Mingqian He, Yongliang Shen, Wenqi Zhang, Qiuying Peng, Jun Wang, Weiming Lu

Generating step-by-step “chain-of-thought” rationales has proven effective for improving the performance of large language models on complex reasoning tasks. However, applying such techniques to structured tasks, such as text-to-SQL, remains largely unexplored. In this paper, we introduce Self-Taught Reasoner for text-to-SQL (STaR-SQL), a novel approach that reframes SQL query generation as a reasoning-driven process. Our method prompts the LLM to produce detailed reasoning steps for SQL queries and fine-tunes it on rationales that lead to correct outcomes. Unlike traditional methods, STaR-SQL dedicates additional test-time computation to reasoning, thereby positioning LLMs as spontaneous reasoners rather than mere prompt-based agents. To further scale the inference process, we incorporate an outcome-supervised reward model (ORM) as a verifier, which enhances SQL query accuracy. Experimental results on the challenging Spider benchmark demonstrate that STaR-SQL significantly improves text-to-SQL performance, achieving an execution accuracy of 86.6\%. This surpasses a few-shot baseline by 31.6\% and a baseline fine-tuned to predict answers directly by 18.0\%. Additionally, STaR-SQL outperforms agent-like prompting methods that leverage more powerful yet closed-source models such as GPT-4. These findings underscore the potential of reasoning-augmented training for structured tasks and open the door to extending self-improving reasoning models to text-to-SQL generation and beyond.

Fairness Beyond Performance: Investigating Reliability Disparities Across Groups in Legal NLP

Santosh T.Y.S.S, Irtiza Chowdhury

Fairness in NLP must extend beyond performance parity to encompass equitable reliability across groups. This study exposes a critical blind spot: models often make less reliable or overconfident predictions for marginalized groups, even when overall performance appears fair. Using the FairLex benchmark as a case study in legal NLP, we systematically evaluate both performance and reliability dispari- ties across demographic, regional, and legal attributes spanning four jurisdictions. We show that domain-specific pre-training consistently improves both performance and reliability, especially for underrepresented groups. However, common bias mitigation methods frequently worsen reliability disparities, revealing a trade-off not captured by performance metrics alone. Our results call for a rethinking of fairness in high-stakes NLP: To ensure equitable treatment, models must not only be accurate, but also reliably self-aware across all groups.

Graph Counselor: Adaptive Graph Exploration via Multi-Agent Synergy to Enhance LLM Reasoning

Junqi Gao, Xiang Zou, Ying Ai, Dong Li, Yichen Niu, Biqing Qi, Jianxing Liu

Graph Retrieval Augmented Generation (GraphRAG) effectively enhances external knowledge integration capabilities by explicitly modeling knowledge relationships, thereby improving the factual accuracy and generation quality of Large Language Models (LLMs) in specialized domains. However, existing methods suffer from two inherent limitations: 1) Inefficient Information Aggregation: They rely on a single agent and fixed iterative patterns, making it difficult to adaptively capture multi-level textual, structural, and degree information within graph data. 2) Rigid Reasoning Mechanism: They employ preset reasoning schemes, which cannot dynamically adjust reasoning depth nor achieve precise semantic correction. To overcome these limitations, we propose Graph Counselor, an GraphRAG method based on multi-agent collaboration. This method uses the Adaptive Graph Information Extraction Module (AGIEM), where Planning, Thought, and Execution Agents work together to precisely model complex graph structures and dynamically adjust information extraction strategies, addressing the challenges of multi-level dependency modeling and adaptive reasoning depth. Additionally, the Self-Reflection with Multiple Perspectives (SR) module improves the accuracy and semantic consistency of reasoning results through self-reflection and backward reasoning mechanisms. Experiments demonstrate that Graph Counselor outperforms existing methods in multiple graph reasoning tasks, exhibiting higher reasoning accuracy and generalization ability. Our code is available at https://github.com/gjq100/Graph-Counselor.git.

Math Neurosurgery: Isolating Language Models' Math Reasoning Abilities Using Only Forward Passes

Bryan R Christ, Zachary Gottesman, Jonathan Kropko, Thomas Hartvigsen

Math reasoning is an active area of Large Language Model (LLM) research because it is a hallmark of artificial intelligence and has implications in several domains, including math education. However, few works have explored how math reasoning is encoded within LLM parameters and if it is a skill that can be isolated within models. Doing so could allow targeted intervention to improve math performance without altering non-math behavior and foster understanding of how models encode math reasoning. We introduce Math Neurosurgery (MathNeuro), a computationally efficient method we use to isolate math-specific parameters in LLMs using only forward passes. MathNeuro builds on existing work by using weights and activations to calculate parameter importance, but isolates math-specific parameters by filtering out those important for general language tasks. Through pruning parameters MathNeuro identifies, we delete a LLM's math reasoning ability without significantly impacting its general language ability. Scaling the identified parameters by a small constant improves a pretrained or instruction-tuned LLM's performance by 4-17% on GSM8K and 5-35% on MATH while leaving non-math behavior unaltered. MathNeuro is also data efficient: most of its effectiveness holds when identifying math-specific parameters using a single sample. MathNeuro highlights the potential for future work to intervene on math-specific parameters.

\mathtt{GeLLM^3O}: Generalizing Large Language Models for Multi-property Molecule Optimization

Vishal Dey, Xiao Hu, Xia Ning

Despite recent advancements, most computational methods for molecule optimization are constrained to single- or double-property optimization tasks and suffer from poor scalability and generalizability to novel optimization tasks. Meanwhile, Large Language Models (LLMs) demonstrate remarkable out-of-domain generalizability to novel tasks. To demonstrate LLMs' potential for molecule optimization, we introduce \mathtt{MuMOInstruct}, the first high-quality instruction-tuning dataset specifically focused on multi-property molecule optimization tasks. Leveraging \mathtt{MuMOInstruct}, we develop \mathtt{GeLLM^3O}s, a series of instruction-tuned LLMs for molecule optimization. Extensive evaluations across 5 in-domain and 5 out-of-domain tasks demonstrate that \mathtt{GeLLM^3O}s consistently outperform state-of-the-art baselines. \mathtt{GeLLM^3O}s also exhibit outstanding zero-shot generalization to unseen tasks, significantly outperforming powerful closed-source LLMs. Such strong generalizability demonstrates the tremendous potential of \mathtt{GeLLM^3O}s as foundational models for molecule optimization, thereby tackling novel optimization tasks without resource-intensive retraining. \mathtt{MuMOInstruct} and code are accessible through https://github.com/ninglab/GeLLMO.

Follow-up Question Generation For Enhanced Patient-Provider Conversations

Joseph Gatto, Parker Seegmiller, Timothy E. Burdick, Inas S. Khayal, Sarah DeLozier, Sarah Masud Preum

Follow-up question generation is an essential feature of dialogue systems as it can reduce conversational ambiguity and enhance modeling complex interactions. Conversational contexts often pose core NLP challenges such as (i) extracting relevant information buried in fragmented data sources, and (ii) modeling parallel thought processes. These two challenges occur frequently in medical dialogue as a doctor asks questions based not only on patient utterances but also their prior EHR data and current diagnostic hypotheses. Asking medical questions in asynchronous conversations compounds these issues as doctors can only rely on static EHR information to motivate follow-up questions. To address these challenges, we introduce FollowupQ, a novel framework for enhancing asynchronous medical conversation. FollowupQ is a multi-agent framework that processes patient messages and EHR data to generate personalized follow-up questions, clarifying patient-reported medical conditions. FollowupQ reduces requisite provider follow-up communications by 34%. It also improves performance by 17% and 5% on real and synthetic data, respectively. We also release the first public dataset of asynchronous medical messages with linked EHR data alongside 2,300 follow-up questions written by clinical experts for the wider NLP research community.

Efficient Ensemble for Fine-tuning Language Models on Multiple Datasets

Dongyue Li, Ziniu Zhang, Lu Wang, Hongyang R. Zhang

This paper develops an ensemble method for fine-tuning a language model to multiple datasets. Existing methods, such as quantized LoRA (QLoRA), are efficient when adapting to a single dataset. When training on multiple datasets of different tasks, a common setup in practice, it remains unclear how to design an efficient adaptation for fine-tuning language models. We propose to use an ensemble of multiple smaller adapters instead of a single adapter per task. We design an efficient algorithm that partitions n datasets into m groups, where m is typically much smaller than n in practice, and train one adapter for each group before taking a weighted combination to form the ensemble. The algorithm leverages a first-order approximation property of low-rank adaptation to quickly obtain the fine-tuning performances of dataset combinations since methods like LoRA stay close to the base model. Hence, we use the gradients of the base model to estimate its behavior during fine-tuning. Empirically, this approximation holds with less than 1% error on models with up to 34 billion parameters, leading to an estimation of true fine-tuning performances under 5% error while speeding up computation compared to base fine-tuning by 105 times. When applied to fine-tune Llama and GPT models on ten text classification tasks, our approach provides up to 10% higher average test accuracy over QLoRA, with only 9% more FLOPs. On a Llama model with 34 billion parameters, an ensemble of QLoRA increases test accuracy by 3% compared to QLoRA, with only 8% more FLOPs.

Embedding-Converter: A Unified Framework for Cross-Model Embedding Transformation

Jinsung Yoon, Sercan O Arik

Embedding models play a crucial role in machine learning. However, the continuous development of new models presents a major challenge: migrating to a potentially superior model often requires the computationally expensive process of re-embedding entire datasets—without any guarantee of performance improvement. This paper presents Embedding-Converter, a novel framework for efficiently transforming embeddings between different models, thus avoiding costly `re-embedding'. The proposed approach achieves 100 times faster and cheaper computations in real-world applications. Experiments show that Embedding-Converter not only streamlines transitions to new models, but can also improve upon the source model's performance, approaching that of the target model. This facilitates efficient evaluation and broader adoption of new embedding models by significantly reducing the overhead of model switching. Furthermore, Embedding-Converter addresses latency limitations by enabling the use of smaller models for online tasks while still benefiting from the performance of larger models offline. By promoting the release of converters alongside new embedding models, Embedding-Converter fosters a more dynamic and accessible ecosystem for embedding model development and deployment.

Efficient Knowledge Editing via Minimal Precomputation

Akshat Gupta, Maochuan Lu, Thomas Hartvigsen, Gopala Anumanchipalli

Knowledge editing methods like MEMIT are able to make data and compute efficient updates of factual knowledge by using a single sentence to update facts and their consequences. However, what is often overlooked is a ``precomputation step'', which requires a one-time but significant computational cost. The authors of MEMIT \cite{MEMIT} originally precompute approximately 44 million hidden vectors per edited layer, which requires a forward pass over 44 million tokens. For GPT-J (6B), this precomputation step takes 36 hours on a single GPU, while it takes approximately 40 hours for Llama2-7B. Additionally, this precomputation time grows with model size. In this paper, we show that this excessive computational cost is unnecessary. Knowledge editing using MEMIT and related methods, such as ROME and EMMET, can be performed by pre-computing a very small portion of the 44 million hidden vectors. We first present the theoretical minimum number of hidden vector precomputation required for solutions of these editing methods to exist. We then empirically show that knowledge editing using these methods can be done by pre-computing significantly fewer hidden vectors. Specifically, we show that the precomputation step can be done with less than 0.3% of the originally stipulated number of hidden vectors. This saves a significant amount of precomputation time and allows users to begin editing new models within a few minutes.

Response Wide Shut:Surprising Observations in Basic Vision Language Model Capabilities

Shivam Chandhok, Wan-Cyuan Fan, Vered Shwartz, Vineeth N. Balasubramanian, Leonid Sigal

Vision-language Models (VLMs) have emerged as general-purpose tools for addressing a variety of complex computer vision problems. Such models have been shown to be highly capable, but, at the same time, lacking some basic visual understanding skills. In this paper, we set out to understand the limitations of SoTA VLMs on fundamental visual tasks (object classification, spatial understanding, and ability to delineate individual object instances through counting), by constructing a series of tests that probe which components of design, specifically, may be lacking. Importantly, we go significantly beyond the current benchmarks, which simply measure the final performance of VLM response, by also comparing and contrasting it to the performance of probes trained directly on features obtained from the visual encoder, intermediate vision-language projection and LLM-decoder output. In doing so, we uncover shortcomings in VLMs and make a number of important observations about their capabilities, robustness and how they process visual information. We hope our insights will guide progress in further improving VLMs.

Balancing the Budget: Understanding Trade-offs Between Supervised and Preference-Based Finetuning

Mohit Raghavendra, Junmo Kang, Alan Ritter

Post-training of Large Language Models often involves a pipeline of Supervised Finetuning (SFT) followed by Preference Finetuning (PFT) using methods like Direct Preference Optimization. Both stages require annotated data that are very different in structure and costs. We study how to optimally allocate a fixed training data budget between the two stages, through extensive experiments spanning four diverse tasks, multiple model sizes and various data annotation costs. Our findings reveal that just SFT on the base model dominates performance in low-data regimes (<1,000 annotated examples). With larger data-budgets, we observe that a combination of SFT and PFT, often with increasing portions allocated towards preference data yields optimal performance. However, completely eliminating SFT and running PFT directly on the base model yields suboptimal performance, described as the cold start problem on tasks like mathematics. We observe that this is due to the distribution shift arising from using DPO directly on the base model to elicit step-by-step reasoning. This limitation can be effectively addressed by allocating even a small portion (<10\%) of the budget to SFT first, resulting in performance improvements of 15-20\% on analytical benchmarks like GSM8k. These results provide actionable insights for researchers and practitioners optimizing model development under budget constraints, where high-quality data curation often represents a significant portion of the total costs of model development.

MindRef: Mimicking Human Memory for Hierarchical Reference Retrieval with Fine-Grained Location Awareness

Ye Wang, Xinrun Xu, Zhiming Ding

When completing knowledge-intensive tasks, humans sometimes need an answer and a corresponding reference passage for auxiliary reading. Previous methods required obtaining pre-segmented article chunks through additional retrieval models. This paper explores leveraging the parameterized knowledge stored during the pre-training phase of large language models (LLMs) to recall reference passage from any starting position independently. We propose a two-stage framework that simulates the scenario of humans recalling easily forgotten references. Initially, the LLM is prompted to recall document title identifiers to obtain a coarse-grained document set. Then, based on the acquired coarse-grained document set, it recalls fine-grained passage. In the two-stage recall process, we use constrained decoding to ensure that content outside of the stored documents is not generated. To increase speed, we only recall a short prefix in the second stage, and then locate its position to retrieve a complete passage. Experiments on KILT knowledge-sensitive tasks have verified that LLMs can independently recall reference passage locations in various task forms, and the obtained reference significantly assists downstream tasks.

HumT DumT: Measuring and controlling human-like language in LLMs

Myra Cheng, Sunny Yu, Dan Jurafsky

Should LLMs generate language that makes them seem human? Human-like language might improve user experience, but might also lead to deception, overreliance, and stereotyping. Assessing these potential impacts requires a systematic way to measure human-like tone in LLM outputs. We introduce HumT and SocioT, metrics for human-like tone and other dimensions of social perceptions in text data based on relative probabilities from an LLM. By measuring HumT across preference and usage datasets, we find that users prefer less human-like outputs from LLMs in many contexts. HumT also offers insights into the perceptions and impacts of anthropomorphism: human-like LLM outputs are highly correlated with warmth, social closeness, femininity, and low status, which are closely linked to the aforementioned harms. We introduce DumT, a method using HumT to systematically control and reduce the degree of human-like tone while preserving model performance. DumT offers a practical approach for mitigating risks associated with anthropomorphic language generation.

ChatBench: From Static Benchmarks to Human-AI Evaluation

Serina Chang, Ashton Anderson, Jake M. Hofman

With the rapid adoption of LLM-based chat-bots, there is a pressing need to evaluate what humans and LLMs can achieve together. However, standard benchmarks, such as MMLU, measure LLM capabilities in isolation (i.e., “AI-alone”). Here, we design and conduct a user study to convert MMLU questions into user-AI conversations, by seeding the user with the question and having them carry out a conversation with the LLM to answer their question. We release ChatBench, a new dataset with AI-alone, user-alone, and user-AI data for 396 questions and two LLMs, including 144K answers and 7,336 user-AI conversations. We find that AI-alone accuracy fails to predict user-AI accuracy, with significant differences across multiple subjects (math, physics, and moral reasoning), and we analyze the user-AI conversations to provide insight into how they diverge from AI-alone benchmarks. Finally, we show that fine-tuning a user simulator on a subset of ChatBench improves its ability to estimate user-AI accuracies, increasing correlation on held-out questions by more than 20 points, creating possibilities for scaling interactive evaluation.

Attention Speaks Volumes: Localizing and Mitigating Bias in Language Models

Rishabh Adiga, Besmira Nushi, Varun Chandrasekaran

We believe that analyzing attention is crucial for understanding bias in large language models (LLMs); in ambiguous comparative prompting frameworks, it provides insight into how the LLM distributes its focus across different entities, and how this contributes to biased decisions. To this end, we first introduce a metric to quantify the "entity preference" of an LLM. We then propose \textbf{ATLAS}, a technique to localize bias to specific layers of the LLM by analyzing attention scores and then reduce bias by scaling attention in these biased layers. To evaluate our method, we conduct extensive experiments across 3 datasets, 4 models, and 4 baseline approaches. Our experiments demonstrate that bias is concentrated in the later layers, typically around the last third. We also show how \textbf{ATLAS} effectively mitigates bias through targeted interventions without compromising downstream performance and an average increase of only 0.34% in perplexity when the intervention is applied. We see an average improvement of 0.28 points in the bias score across all the datasets.

The Efficiency vs. Accuracy Trade-off: Optimizing RAG-Enhanced LLM Recommender Systems Using Multi-Head Early Exit

Huixue Zhou, Hengrui Gu, Zaifu Zhan, Xi Liu, Kaixiong Zhou, Yongkang Xiao, Mingfu Liang, Srinivas Prasad Govindan, Piyush Chawla, Jiyan Yang, Xiangfei Meng, Huayu Li, Buyun Zhang, Liang Luo, Wen-Yen Chen, Yiping Han, Bo Long, Rui Zhang, Tianlong Chen

The deployment of Large Language Models (LLMs) in recommender systems for Click-Through Rate (CTR) prediction requires a careful balance between computational efficiency and predictive accuracy. This paper introduces OptiRAG-Rec, a comprehensive framework that integrates Retrieval-Augmented Generation (RAG) with a novel multi-head early exit architecture to address both challenges. By leveraging Graph Convolutional Networks (GCNs) as efficient retrieval mechanisms, the framework significantly reduces data retrieval times while maintaining high model performance. Additionally, the multi-head early exit strategy dynamically terminates inference based on real-time predictive confidence assessments, enhancing responsiveness without sacrificing accuracy. Experimental results demonstrate that OptiRAG-Rec reduces computation time while preserving the precision required for reliable recommendations, establishing a new benchmark for efficient and accurate LLM deployment in recommendation.

Unraveling LoRA Interference: Orthogonal Subspaces for Robust Model Merging

Haobo Zhang, Jiayu Zhou

Fine-tuning large language models (LMs) for individual tasks yields strong performance but is expensive for deployment and storage. Recent works explore model merging to combine multiple task-specific models into a single multi-task model without additional training. However, existing merging methods often fail for models fine-tuned with low-rank adaptation (LoRA), due to significant performance degradation. In this paper, we show that this issue arises from a previously overlooked interplay between model parameters and data distributions. We propose **O**rthogonal **S**ubspaces for **R**obust model **M**erging (**OSRM**) to constrain the LoRA subspace *prior* to fine-tuning, ensuring that updates relevant to one task do not adversely shift outputs for others. Our approach can seamlessly integrate with most existing merging algorithms, reducing the unintended interference among tasks. Extensive experiments on eight datasets, tested with three widely used LMs and two large LMs, demonstrate that our method not only boosts merging performance but also preserves single-task accuracy. Furthermore, our approach exhibits greater robustness to the hyperparameters of merging. These results highlight the importance of data-parameter interaction in model merging and offer a plug-and-play solution for merging LoRA models.

CSTree-SRI: Introspection-Driven Cognitive Semantic Tree for Multi-Turn Question Answering over Extra-Long Contexts

Zhaowen Wang, Xiang Wei, Kangshao Du, Yiting Zhang, Libo Qin, Yingjie Xia, Li Kuang

Large Language Models (LLMs) have achieved remarkable success in natural language processing (NLP), particularly in single-turn question answering (QA) on short-text. However, their performance significantly declines when applied to multi-turn QA over extra-long context (ELC), as they struggle to capture the logical correlations across multiple chunks of ELC and maintain the coherence of multi-turn Questions. To address the challenges, we propose the CSTree-SRI framework (Cognitive Semantic Tree through Summarization, Retrieval, and Introspection). CSTree-SRI dynamically constructs the CSTree to preserve logical coherence within ELC through hierarchical synthesis and introspective validation. Then a logic-driven traversal strategy on CSTree is designed to provide efficient information retrieval for question answering. Additionally, we construct a suite of multi-turn QA datasets and an evaluation benchmark tailored for ELC tasks, and comprehensive experiments demonstrate the framework's superiority in addressing the challenges of multi-turn QA over ELC.

RATIONALYST: Pre-training Process-Supervision for Improving Reasoning

Dongwei Jiang, Guoxuan Wang, Yining Lu, Andrew Wang, Jingyu Zhang, Chuyu Liu, Benjamin Van Durme, Daniel Khashabi

The reasoning steps generated by LLMs might be incomplete, as they mimic logical leaps common in everyday communication found in their pre-training data: underlying rationales are frequently left implicit (unstated). To address this challenge, we introduce RATIONALYST, a model for process-supervision of reasoning based on pre-training on a vast collection of rationale annotations extracted from unlabeled data. We extract 79k rationales from web-scale unlabelled dataset (the Pile) and a combination of reasoning datasets with minimal human intervention. This web-scale pre-training for reasoning allows RATIONALYST to consistently generalize across diverse reasoning tasks, including mathematical, commonsense, scientific, and logical reasoning. Fine-tuned from LLaMa-3-8B, RATIONALYST improves the accuracy of reasoning by an average of 3.9% on 7 representative reasoning benchmarks. It also demonstrates superior performance compared to significantly larger verifiers like GPT-4 and similarly sized models fine-tuned on matching training sets.

Advancing SMoE for Continuous Domain Adaptation of MLLMs: Adaptive Router and Domain-Specific Loss

Liang Zhang, Ziyao Lu, Fandong Meng, Hui Li, Jie Zhou, Jinsong Su

Recent studies have explored Continual Instruction Tuning (CIT) in Multimodal Large Language Models (MLLMs), with a primary focus on Task-incremental CIT, where MLLMs are required to continuously acquire new tasks. However, the more practical and challenging Domain-incremental CIT, focused on the continual adaptation of MLLMs to new domains, remains underexplored. In this paper, we propose a new Sparse Mixture of Expert (SMoE) based method for domain-incremental CIT in MLLMs. During training, we learn a domain-specific SMoE module for each new domain in every FFN sub-layer of MLLMs, preventing catastrophic forgetting caused by inter-domain conflicts. Moreover, we equip the SMoE module with a domain-specific autoregressive loss (DSAL), which is used to identify the most suitable SMoE module for processing each test instruction during inference. To further enhance the SMoE module's ability to learn domain knowledge, we design an adaptive threshold-based router (AT-Router) that allocates computing resources (experts) to instruction tokens based on their importance. Finally, we establish a new benchmark to evaluate the efficacy of our method and advance future research. Extensive experiments show that our method consistently outperforms all competitive baselines.

Mitigating Media Bias through Multi-document Events Reasoning in LLMs

Yuanyuan Lei, Ruihong Huang

Media outlets are becoming more partisan and polarized nowadays. Most previous work focused on detecting media bias. In this paper, we aim to mitigate media bias by generating a neutralized summary given multiple articles presenting different ideological views. Motivated by the critical role of events and event relations in media bias detection, we propose to increase awareness of bias in LLMs via multi-document events reasoning and use a multi-document event relation graph to guide the summarization process. This graph contains rich event information useful to reveal bias: four common types of in-doc event relations to reflect content framing bias, cross-doc event coreference relation to reveal content selection bias, and event-level moral opinions to highlight opinionated framing bias. We further develop two strategies to incorporate the multi-document event relation graph for neutralized summarization. Firstly, we convert a graph into natural language descriptions and feed the textualized graph into LLMs as a part of a hard text prompt. Secondly, we encode the graph with graph attention network and insert the graph embedding into LLMs as a soft prompt. Both automatic evaluation and human evaluation confirm that our approach effectively mitigates both lexical and informational media bias, and meanwhile improves content preservation.

RoCoFT: Efficient Finetuning of Large Language Models with Row-Column Updates

Md Kowsher, Tara Esmaeilbeig, Chun-Nam Yu, Chen Chen, Mojtaba Soltanalian, Niloofar Yousefi

We propose Row-Column Fine-Tuning(RoCoFT), a parameter-efficient fine-tuning method for large language models based on updating only a few rows and columns of the weight matrices in transformers. Through extensive experiments with medium-sized LMs like RoBERTa and DeBERTa, and larger LMs like Bloom-7B, Llama2-7B, and Llama2-13B, we show that our method gives comparable or better accuracies than state-of-the-art Parameter-Efficient Finetuning methods while also being more memory and computation-efficient. We also study the reason behind the effectiveness of our method with tools from neural tangent kernel theory. We empirically demonstrate that our kernel, constructed using a restricted set of row and column parameters, is numerically close to the full-parameter kernel and gives comparable classification performance. Ablation studies are conducted to investigate the impact of different algorithmic choices, including the robustness of RoCoFT to any selection of rows and columns, as well as the optimal rank for the effective implementation of our method.

VideoVista2: 360° Horizons-Bridging Cultures, Languages, and Domains in Video Comprehension

Xinyu Chen, yunxin li, Haoyuan Shi, Baotian Hu, Wenhan Luo, Yaowei Wang, Min Zhang

Assessing the video comprehension capabilities of multimodal AI systems can effectively measure their understanding and reasoning abilities. Most video evaluation benchmarks are limited to a single language, typically English, and predominantly feature videos rooted in Western cultural contexts. In this paper, we present **VideoVista-CulturalLingo**, the first video evaluation benchmark designed to bridge cultural, linguistic, and domain divide in video comprehension. Our work differs from existing benchmarks in the following ways: 1) **Cultural diversity**, incorporating cultures from China, North America, and Europe; 2) **Multi-linguistics**, with questions presented in Chinese and English—two of the most widely spoken languages; and 3) **Broad domain**, featuring videos sourced from hundreds of human-created domains. VideoVista-CulturalLingo contains 1,389 videos and 3,134 QA pairs, and we have evaluated 24 recent open-source or proprietary video large models. From the experiment results, we observe that: 1) Existing models perform worse on Chinese-centric questions than Western-centric ones, particularly those related to Chinese history; 2) Current open-source models still exhibit limitations in temporal understanding, especially in the Event Localization task, achieving a maximum score of only 45.2\%; 3) Mainstream models demonstrate strong performance in general scientific questions, while open-source models demonstrate weak performance in mathematics.

FOCUS: Evaluating Pre-trained Vision-Language Models on Underspecification Reasoning

Kankan Zhou, Eason Lai, Kyriakos Mouratidis, Jing Jiang

Humans possess a remarkable ability to interpret underspecified ambiguous statements by inferring their meanings from contexts such as visual inputs. This ability, however, may not be as developed in recent pre-trained vision-language models(VLMs). In this paper, we introduce a novel probing dataset called FOCUS to evaluate whether state-of-the-art VLMs have this ability. FOCUS consists of underspecified sentences paired with image contexts and carefully designed probing questions. Our experiments reveal that VLMs still fall short in handling underspecification even when visual inputs that can help resolve the ambiguities are available. To further support research in underspecification, FOCUS will be released for public use. We hope this dataset will inspire further research on the reasoning and contextual understanding capabilities of VLMs.

Enhancing Input-Label Mapping in In-Context Learning with Contrastive Decoding

Keqin Peng, Liang Ding, Yuanxin Ouyang, Meng Fang, Yancheng Yuan, Dacheng Tao

Large language models (LLMs) excel at a range of tasks through in-context learning (ICL), where only a few task examples guide their predictions. However, prior research highlights that LLMs often overlook input-label mapping information in ICL, relying more on their pre-trained knowledge. To address this issue, we introduce In-Context Contrastive Decoding (ICCD), a novel method that emphasizes input-label mapping by contrasting the output distributions between positive and negative in-context examples. Experiments on 7 natural language understanding (NLU) tasks show that our ICCD method brings consistent and significant improvement (up to +1.8 improvement on average) upon 6 different scales of LLMs without requiring additional training. Our approach is versatile, enhancing performance with various demonstration selection methods, demonstrating its broad applicability and effectiveness. The code and scripts are released at https://github.com/Romainpkq/CD_ICL.

Value Residual Learning

Zhanchao Zhou, Tianyi Wu, Zhiyun Jiang, Fares Obeid, Zhenzhong Lan

While Transformer models have achieved remarkable success in various domains, the effectiveness of information propagation through deep networks remains a critical challenge. Standard hidden state residuals often fail to adequately preserve initial token-level information in deeper layers. This paper introduces ResFormer, a novel architecture that enhances information flow by incorporating value residual connections in addition to hidden state residuals. And a variant is SVFormer, where all layers share the first layer's value embedding. Comprehensive empirical evidence demonstrates ResFormer achieves equivalent validation loss with 16.11\% fewer model parameters and 20.3\% less training data compared to Transformer, while maintaining similar memory usage and computational cost. Besides, SVFormer reduces KV cache size by nearly half with only a small performance penalty and can be integrated with other KV-efficient methods, yielding further reductions in KV cache, with performance influenced by sequence length and cumulative learning rate.

ERU-KG: Efficient Reference-aligned Unsupervised Keyphrase Generation

Lam Thanh Do, Aaditya Bodke, Pritom Saha Akash, Kevin Chen-Chuan Chang

Unsupervised keyphrase prediction has gained growing interest in recent years. However, existing methods typically rely on heuristically defined importance scores, which may lead to inaccurate informativeness estimation. In addition, they lack consideration for time efficiency. To solve these problems, we propose ERU-KG, an unsupervised keyphrase generation (UKG) model that consists of an informativeness and a phraseness module. The former estimates the relevance of keyphrase candidates, while the latter generate those candidates. The informativeness module innovates by learning to model informativeness through references (e.g., queries, citation contexts, and titles) and at the term-level, thereby 1) capturing how the key concepts of documents are perceived in different contexts and 2) estimating informativeness of phrases more efficiently by aggregating term informativeness, removing the need for explicit modeling of the candidates. ERU-KG demonstrates its effectiveness on keyphrase generation benchmarks by outperforming unsupervised baselines and achieving on average 89% of the performance of a supervised model for top 10 predictions. Additionally, to highlight its practical utility, we evaluate the model on text retrieval tasks and show that keyphrases generated by ERU-KG are effective when employed as query and document expansions. Furthermore, inference speed tests reveal that ERU-KG is the fastest among baselines of similar model sizes. Finally, our proposed model can switch between keyphrase generation and extraction by adjusting hyperparameters, catering to diverse application requirements.

Know Your Mistakes: Towards Preventing Overreliance on Task-Oriented Conversational AI Through Accountability Modeling

Suvodip Dey, Yi-Jyun Sun, Gokhan Tur, Dilek Hakkani-Tür

Recent LLMs have enabled significant advancements for conversational agents. However, they are also well known to hallucinate, producing responses that seem plausible but are factually incorrect. On the other hand, users tend to over-rely on LLM-based AI agents, accepting AI’s suggestion even when it is wrong. Adding positive friction, such as explanations or getting user confirmations, has been proposed as a mitigation in AI-supported decision-making systems. In this paper, we propose an accountability model for LLM-based task-oriented dialogue agents to address user overreliance via friction turns in cases of model uncertainty and errors associated with dialogue state tracking (DST). The accountability model is an augmented LLM with an additional accountability head that functions as a binary classifier to predict the relevant slots of the dialogue state mentioned in the conversation. We perform our experiments with multiple backbone LLMs on two established benchmarks (MultiWOZ and Snips). Our empirical findings demonstrate that the proposed approach not only enables reliable estimation of AI agent errors but also guides the decoder in generating more accurate actions. We observe around 3% absolute improvement in joint goal accuracy (JGA) of DST output by incorporating accountability heads into modern LLMs. Self-correcting the detected errors further increases the JGA from 67.13 to 70.51, achieving state-of-the-art DST performance. Finally, we show that error correction through user confirmations (friction turn) achieves a similar performance gain, highlighting its potential to reduce user overreliance.

Misattribution Matters: Quantifying Unfairness in Authorship Attribution

Pegah Alipoormolabashi, Ajay Patel, Niranjan Balasubramanian

Authorship misattribution can have profound consequences in real life. In forensic settings simply being considered as one of the potential authors of an evidential piece of text or communication can result in undesirable scrutiny. This raises a fairness question: Is every author in the candidate pool at equal risk of misattribution? Standard evaluation measures for authorship attribution systems do not explicitly account for this notion of fairness. We introduce a simple measure, Misattribution Unfairness Index (MAUI_k), which is based on how often authors are ranked in the top k for texts they did not write. Using this measure we quantify the unfairness of five models on two different datasets. All models exhibit high levels of unfairness with increased risks for some authors. Furthermore, we find that this unfairness relates to how the models embed the authors as vectors in the latent search space. In particular, we observe that the risk of misattribution is higher for authors closer to the centroid (or center) of the embedded authors in the haystack. These results indicate the potential for harm and the need for communicating with and calibrating end users on misattribution risk when building and providing such models for downstream use.

Class Distillation with Mahalanobis Contrast: An Efficient Training Paradigm for Pragmatic Language Understanding Tasks

Chenlu Wang, Weimin Lyu, Ritwik Banerjee

Detecting deviant language such as sexism, or nuanced language such as metaphors or sarcasm, is crucial for enhancing the safety, clarity, and interpretation of social interactions. While existing classifiers deliver strong results on these tasks, they often come with significant computational cost and high data demands. In this work, we propose \textbf{Cla}ss \textbf{D}istillation (ClaD), a novel training paradigm that targets the core challenge: distilling a small, well-defined target class from a highly diverse and heterogeneous background. ClaD integrates two key innovations: (i) a loss function informed by the structural properties of class distributions, based on Mahalanobis distance, and (ii) an interpretable decision algorithm optimized for class separation. Across three benchmark detection tasks -- sexism, metaphor, and sarcasm -- ClaD outperforms competitive baselines, and even with smaller language models and orders of magnitude fewer parameters, achieves performance comparable to several large language models. These results demonstrate ClaD as an efficient tool for pragmatic language understanding tasks that require gleaning a small target class from a larger heterogeneous background.

SEE: Strategic Exploration and Exploitation for Cohesive In-Context Prompt Optimization

Wendi Cui, Jiaxin Zhang, Zhuohang Li, Hao Sun, Damien Lopez, Kamalika Das, Bradley A. Malin, Sricharan Kumar

Designing optimal prompts for Large Language Models (LLMs) is a complex and resource-intensive task, often requiring substantial human expertise. Existing approaches typically separate the optimization of prompt instructions and in-context learning examples, leading to incohesive, suboptimal results. To overcome this limitation, we propose a novel Cohesive In-Context Prompt Optimization framework that refines both prompt instructions and examples. In our formulation, coherence refers to the degree to which instructions and examples work synergistically to improve task performance—emerging as a byproduct of performance-driven optimization. However, formulating such an optimization in the discrete and high-dimensional space of natural language poses significant challenges in both convergence and computational efficiency. To address these issues, we introduce SEE, a scalable and efficient prompt optimization framework that adopts metaheuristic optimization principles and strategically balances exploration and exploitation to enhance optimization performance and achieve efficient convergence. SEE features a quad-phased design that alternates between global traversal (exploration) and local optimization (exploitation) and adaptively chooses LLM operators during the optimization process. We have conducted a comprehensive evaluation across 35 benchmark tasks, and SEE significantly outperforms state-of-the-art baseline methods by a large margin, achieving an average performance gain of **13.94** while reducing computational costs by **58.67%**.

The Task Shield: Enforcing Task Alignment to Defend Against Indirect Prompt Injection in LLM Agents

Feiran Jia, Tong Wu, Xin Qin, Anna Squicciarini

Large Language Model (LLM) agents are increasingly being deployed as conversational assistants capable of performing complex real-world tasks through tool integration. This enhanced ability to interact with external systems and process various data sources, while powerful, introduces significant security vulnerabilities. In particular, indirect prompt injection attacks pose a critical threat, where malicious instructions embedded within external data sources can manipulate agents to deviate from user intentions. While existing defenses show promise, they struggle to maintain robust security while preserving task functionality. We propose a novel and orthogonal perspective that reframes agent security from preventing harmful actions to ensuring task alignment, requiring every agent action to serve user objectives. Based on this insight, we develop Task Shield, a test-time defense mechanism that systematically verifies whether each instruction and tool call contributes to user-specified goals. Through experiments on the AgentDojo benchmark, we demonstrate that Task Shield reduces attack success rates (2.07\%) while maintaining high task utility (69.79\%) on GPT-4o, significantly outperforming existing defenses in various real-world scenarios.

TreeCut: A Synthetic Unanswerable Math Word Problem Dataset for LLM Hallucination Evaluation

Jialin Ouyang

Large language models (LLMs) now achieve near-human performance on standard math word problem benchmarks (e.g., GSM8K), yet their true reasoning ability remains disputed. A key concern is that models often produce confident, yet unfounded, answers to unanswerable problems. We introduce TreeCut, a synthetic dataset that systematically generates infinite unanswerable math word problems and their answerable counterparts, by representing each question as a tree and removing chosen necessary conditions. Experiments show TreeCut effectively induce hallucinations in large language models, including GPT-4o and o3-mini, with rates of 64% and 44% in their respective worst-case scenarios under zero-shot setting. Further analysis highlights that deeper or more complex trees, composite item names, and removing necessary condition near the middle of a path all increase the likelihood of hallucinations, underscoring the persistent challenges LLMs face in identifying unanswerable math problems. The dataset generation code and sample data are available at .

Symmetrical Visual Contrastive Optimization: Aligning Vision-Language Models with Minimal Contrastive Images

Shengguang Wu, Fan-Yun Sun, Kaiyue Wen, Nick Haber

Recent studies have shown that Large Vision-Language Models (VLMs) tend to neglect image content and over-rely on language-model priors, resulting in errors in visually grounded tasks and hallucinations. We hypothesize that this issue arises because existing VLMs are not explicitly trained to generate texts that are accurately grounded in fine-grained image details. To enhance visual feedback during VLM training, we propose S-VCO (Symmetrical Visual Contrastive Optimization), a novel finetuning objective that steers the model toward capturing important visual details and aligning them with corresponding text tokens. To further facilitate this detailed alignment, we introduce MVC, a paired image-text dataset built by automatically filtering and augmenting visual counterfactual data to challenge the model with hard contrastive cases involving Minimal Visual Contrasts. Experiments show that our method consistently improves VLM performance across diverse benchmarks covering various abilities and domains, achieving up to a 22% reduction in hallucinations, and significant gains in vision-centric and general tasks. Notably, these improvements become increasingly pronounced in benchmarks with higher visual dependency. In short, S-VCO offers a significant enhancement of VLM's visually-dependent task performance while retaining or even improving the model's general abilities.

CHEER-Ekman: Fine-grained Embodied Emotion Classification

Phan Anh Duong, Cat Luong, Divyesh Bommana, Tianyu Jiang

Emotions manifest through physical experiences and bodily reactions, yet identifying such embodied emotions in text remains understudied. We present an embodied emotion classification dataset, CHEER-Ekman, extending the existing binary embodied emotion dataset with Ekman's six basic emotion categories. Using automatic best-worst scaling with large language models, we achieve performance superior to supervised approaches on our new dataset. Our investigation reveals that simplified prompting instructions and chain-of-thought reasoning significantly improve emotion recognition accuracy, enabling smaller models to achieve competitive performance with larger ones.

Improving Fairness of Large Language Models in Multi-document Summarization

Haoyuan Li, Rui Zhang, Snigdha Chaturvedi

Fairness in multi-document summarization (MDS) is crucial for providing comprehensive views across documents with diverse social attribute values, which can significantly impact decision-making. For example, a summarization system that tends to overrepresent negative reviews of products can mislead customers into disregarding good products. Previous works measure fairness in MDS at two levels: summary-level and corpus-level. While summary-level fairness focuses on individual summaries, corpus-level fairness focuses on a corpus of summaries. Recent methods primarily focus on summary-level fairness. We propose FairPO, a preference tuning method that focuses on both summary-level and corpus-level fairness in MDS. To improve summary-level fairness, we propose to generate preference pairs by perturbing document sets. To improve corpus-level fairness, we propose fairness-aware preference tuning by dynamically adjusting the weights of preference pairs. Our experiments show that FairPO outperforms strong baselines while maintaining the critical qualities of summaries. The code is available at https://github.com/leehaoyuan/coverage_fairness

Quantized Can Still Be Calibrated: A Unified Framework to Calibration in Quantized Large Language Models

Mingyu Zhong, Guanchu Wang, Yu-Neng Chuang, Na Zou

Although weight quantization helps large language models (LLMs) in resource-constrained environments, its influence on the uncertainty calibration remains unexplored. To bridge this gap, we presents a comprehensive investigation of uncertainty calibration for quantized LLMs in this work. Specifically, we propose an analytic method to estimate the upper bound of calibration error (UBCE) for LLMs. Our method separately discusses the calibration error of the model's correct and incorrect predictions, indicating a theoretical improvement of calibration error caused by the weight quantization. Our study demonstrates that quantized models consistently exhibit worse calibration performance than full-precision models, supported by consistent analysis across multiple LLMs and datasets. To address the calibration issues of quantized models, we propose a novel method of post calibration for recovering the calibration performance of quantized models through soft-prompt tuning. Specifically, we inject soft tokens to quantized models after the embedding layers, and optimize these tokens to recover the calibration error caused by the weight quantization. Experimental results on multiple datasets demonstrate our effectiveness in improving the uncertainty calibration of quantized LLMs, facilitating more reliable weight quantization in resource-constrained environments.

Astute RAG: Overcoming Imperfect Retrieval Augmentation and Knowledge Conflicts for Large Language Models

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, Sercan O Arik

Retrieval augmented generation (RAG), while effectively integrating external knowledge to address the inherent limitations of large language models (LLMs), can be hindered by imperfect retrieval that contain irrelevant, misleading, or even malicious information. Previous studies have rarely connected the behavior of RAG through joint analysis, particularly regarding error propagation coming from imperfect retrieval and potential conflicts between LLMs' internal knowledge and external sources. Through comprehensive and controlled analyses under realistic conditions, we find that imperfect retrieval augmentation is inevitable, common, and harmful. We identify the knowledge conflicts between LLM-internal and external knowledge from retrieval as a bottleneck to overcome imperfect retrieval in the post-retrieval stage of RAG. To address this, we propose Astute RAG, a novel RAG approach designed to be resilient to imperfect retrieval augmentation. It adaptively elicits essential information from LLMs' internal knowledge, iteratively consolidates internal and external knowledge with source-awareness, and finalizes the answer according to information reliability. Our experiments with Gemini and Claude demonstrate the superior performance of Astute RAG compared to previous robustness-enhanced RAG approaches. Specifically, Astute RAG is the only RAG method that achieves performance comparable to or even surpassing conventional use of LLMs under the worst-case scenario. Further analysis reveals the effectiveness of Astute RAG in resolving knowledge conflicts, thereby improving the trustworthiness of RAG.

M3GQA: A Multi-Entity Multi-Hop Multi-Setting Graph Question Answering Benchmark

Boci Peng, Yongchao Liu, Xiaohe Bo, Jiaxin Guo, Yun Zhu, Xuanbo Fan, Chuntao Hong, Yan Zhang

Recently, GraphRAG systems have achieved remarkable progress in enhancing the performance and reliability of large language models (LLMs). However, most previous benchmarks are template-based and primarily focus on few-entity queries, which are monotypic and simplistic, failing to offer comprehensive and robust assessments. Besides, the lack of ground-truth reasoning paths also hinders the assessments of different components in GraphRAG systems. To address these limitations, we propose \text{M}^3\text{GQA}, a complex, diverse, and high-quality GraphRAG benchmark focusing on multi-entity queries, with six distinct settings for comprehensive evaluation. In order to construct diverse data with semantically correct ground-truth reasoning paths, we introduce a novel reasoning-driven four-step data construction method, including tree sampling, reasoning path backtracking, query creation, and multi-stage refinement and filtering. Extensive experiments demonstrate that \text{M}^3\text{GQA} effectively reflects the capabilities of GraphRAG methods, offering valuable insights into the model performance and reliability. By pushing the boundaries of current methods, \text{M}^3\text{GQA} establishes a comprehensive, robust, and reliable benchmark for advancing GraphRAG research.

TACLR: A Scalable and Efficient Retrieval-based Method for Industrial Product Attribute Value Identification

Yindu Su, Huike Zou, Lin Sun, Ting Zhang, Haiyang Yang, Chen Li Yu, David Lo, qingheng zhang, Shuguang Han, jufeng chen

Product Attribute Value Identification (PAVI) involves identifying attribute values from product profiles, a key task for improving product search, recommendation, and business analytics on e-commerce platforms. However, existing PAVI methods face critical challenges, such as inferring implicit values, handling out-of-distribution (OOD) values, and producing normalized outputs. To address these limitations, we introduce Taxonomy-Aware Contrastive Learning Retrieval (TACLR), the first retrieval-based method for PAVI. TACLR formulates PAVI as an information retrieval task by encoding product profiles and candidate values into embeddings and retrieving values based on their similarity. It leverages contrastive training with taxonomy-aware hard negative sampling and employs adaptive inference with dynamic thresholds. TACLR offers three key advantages: (1) it effectively handles implicit and OOD values while producing normalized outputs; (2) it scales to thousands of categories, tens of thousands of attributes, and millions of values; and (3) it supports efficient inference for high-load industrial deployment. Extensive experiments on proprietary and public datasets validate the effectiveness and efficiency of TACLR. Further, it has been successfully deployed on the real-world e-commerce platform Xianyu, processing millions of product listings daily with frequently updated, large-scale attribute taxonomies. We release the code to facilitate reproducibility and future research at https://github.com/SuYindu/TACLR.

RePanda: Pandas-powered Tabular Verification and Reasoning

Atoosa Chegini, Keivan Rezaei, Hamid Eghbalzadeh, Soheil Feizi

Fact-checking tabular data is essential for ensuring the accuracy of structured information in domains such as journalism, finance, and scientific research. However, existing methods often rely on black-box models with opaque reasoning. We introduce RePanda, a structured fact verification approach that translates claims into executable pandas queries, enabling interpretable and verifiable reasoning. To train RePanda, we construct PanTabFact, a structured dataset derived from TabFact, where claims are paired with executable queries generated using DeepSeek-Chat and refined through automated error correction. Fine-tuning DeepSeek-coder-7B-instruct-v1.5 on PanTabFact, RePanda achieves 84.09% accuracy on TabFact. To assess Out-of-Distribution (OOD) generalization, we create a dataset named WikiFact from WikiTableQuestions by transforming question-answer pairs into factual claims. Without additional fine-tuning, RePanda achieves 84.72% accuracy on WikiFact, significantly outperforming all other baselines and demonstrating strong OOD robustness. PanTabFact is publically available on HuggingFace at datasets/AtoosaChegini/PanTabFact. Beyond fact verification, RePanda extends to tabular question answering by generating executable queries that retrieve precise answers. To support this, we introduce PanWiki, a dataset mapping WikiTableQuestions to pandas queries. Fine-tuning on PanWiki, RePanda achieves 75.1% accuracy in direct answer retrieval. These results highlight the effectiveness of structured execution-based reasoning for tabular verification and question answering.

Low-Bit Quantization Favors Undertrained LLMs: Scaling Laws for Quantized LLMs with 100T Training Tokens

Xu Ouyang, Tao Ge, Thomas Hartvigsen, Zhisong Zhang, Haitao Mi, Dong Yu

Low-bit quantization improves machine learning model efficiency but surprisingly favors undertrained large language models (LLMs). Larger models or those trained on fewer tokens exhibit less quantization-induced degradation (QiD), while smaller, well-trained models face significant performance losses. To gain deeper insights into this trend, we study over 1500+ quantized LLM checkpoints of various sizes and at different training levels (undertrained or fully trained) in a controlled setting, deriving scaling laws for understanding the relationship between QiD and factors: the number of training tokens, model size and bit width. With our derived scaling laws, we propose a novel perspective that we can use QiD to measure an LLM's training levels and determine the number of training tokens required for fully training LLMs of various sizes. Moreover, we use the scaling laws to predict the quantization performance of different-sized LLMs trained with tokens. Our projection shows that the low-bit quantization performance of future models, which are expected to be trained with over \textcolor{red}{100trillion} tokens, may NOT be desirable. This poses a potential challenge for low-bit quantization in the future and highlights the need for awareness of a model's training level when evaluating low-bit quantization research. To facilitate future research on this problem, we release all the 1500+ quantized checkpoints used in this work at https://huggingface.co/Xu-Ouyang.

Energy Considerations of Large Language Model Inference and Efficiency Optimizations

Jared Fernandez, Clara Na, Vashisth Tiwari, Yonatan Bisk, Sasha Luccioni, Emma Strubell

As large language models (LLMs) scale in size and adoption, their computational and environmental costs continue to rise. Prior benchmarking efforts have primarily focused on latency reduction in idealized settings, often overlooking the diverse real-world inference workloads that shape energy use. In this work, we systematically analyze the energy implications of common inference efficiency optimizations across diverse Natural Language Processing (NLP) and generative Artificial Intelligence (AI) workloads, including conversational AI and code generation. We introduce a modeling approach that approximates real-world LLM workflows through a binning strategy for input-output token distributions and batch size variations. Our empirical analysis spans software frameworks, decoding strategies, GPU architectures, online and offline serving settings, and model parallelism configurations. We show that the effectiveness of inference optimizations is *highly sensitive to workload geometry, software stack, and hardware accelerators*, demonstrating that naive energy estimates based on FLOPs or theoretical GPU utilization significantly underestimate real-world energy consumption. Our findings reveal that the proper application of relevant inference efficiency optimizations can reduce total energy use by up to **73%** from unoptimized baselines. These insights provide a foundation for sustainable LLM deployment and inform energy-efficient design strategies for future AI infrastructure.

BFS-Prover: Scalable Best-First Tree Search for LLM-based Automatic Theorem Proving

Ran Xin, Chenguang Xi, Jie Yang, Feng Chen, Hang Wu, Xia Xiao, Yifan Sun, Shen Zheng, Ming Ding

Recent advancements in large language models (LLMs) have spurred growing interest in automatic theorem proving using Lean4, where effective tree search methods are crucial for navigating the underlying large proof search spaces. While the existing approaches primarily rely on value functions and/or Monte Carlo Tree Search (MCTS), the potential of simpler methods like Best-First Tree Search (BFS) remains underexplored. In this paper, we investigate whether BFS can achieve competitive performance in large-scale theorem proving tasks. We present \texttt{BFS-Prover}, a scalable expert iteration framework, featuring three key innovations. First, we implement strategic data filtering at each expert iteration round, excluding problems solvable via beam search node expansion to focus on harder cases. Second, we improve the sample efficiency of BFS through Direct Preference Optimization (DPO) applied to state-tactic pairs automatically annotated with compiler error feedback, refining the LLM's policy to prioritize productive expansions. Third, we employ length normalization in BFS to encourage exploration of deeper proof paths. \texttt{BFS-Prover} achieves a state-of-the-art score of 72.95 on the MiniF2F test set and therefore challenges the perceived necessity of complex tree search methods, demonstrating that BFS can achieve competitive performance when properly scaled.

LAMB: A Training-Free Method to Enhance the Long-Context Understanding of SSMs via Attention-Guided Token Filtering

Zhifan Ye, Zheng Wang, Kejing Xia, Jihoon Hong, Leshu Li, Lexington Whalen, Cheng Wan, Yonggan Fu, Yingyan Celine Lin, Souvik Kundu

State space models (SSMs) achieve efficient sub-quadratic compute complexity but often exhibit significant performance drops as context length increases. Recent work attributes this deterioration to an exponential decay in hidden-state memory. While token filtering has emerged as a promising remedy, its underlying rationale and limitations remain largely non-understood. In this paper, we first investigate the attention patterns of Mamba to shed light on why token filtering alleviates long-context degradation. Motivated by these findings, we propose LAMB, a training-free, attention-guided token filtering strategy designed to preserve critical tokens during inference. LAMB can boost long-context performance for both pure SSMs and hybrid models, achieving up to an average improvement of 30.35% over state-of-the-art techniques on standard long-context understanding benchmarks. Our analysis and experiments reveal new insights into the interplay between attention, token selection, and memory retention, and are thus expected to inspire broader applications of token filtering in long-sequence modeling.

Can Multi-turn Self-judge Single Agent LMs with Retrieval Solve ICPC Problems?

Md. Tanzib Hosain, Md Kishor Morol

Among the hardest tasks for humans are those found in competitive programming where problems require sophisticated algorithmic thinking, puzzle solving, and the creation of effective code. As a domain to assess language models (LMs), it has not received enough attention, though. This study presents the ICPC benchmark, which consists of 1220 international collegiate programming contest (ICPC) tasks. Each problem includes official analysis, reference code, and sample and high-quality unit and hidden tests. We are able to develop and evaluate a variety of LM inference techniques for competitive programming with these resources. With zero-shot chain-of-thought prompting, we find that o1 only achieves a 19.1\% pass@1 solve rate. With our best inference technique, which combines muti-turn self-judge with reflection and retrieval over episodic information, raises this to 42.2\%. Furthermore, we conduct a new human-in-the-loop investigation to gain a deeper understanding of the remaining difficulties. Surprisingly, we discover that o1 can solve 17 out of 18 problems that were previously unsolvable by any model or technique with just a few specific instructions. A footstep toward LMs with grounded, imaginative, and algorithmic thinking is provided by our quantitative findings and qualitative research.

Do Androids Question Electric Sheep? A Multi-Agent Cognitive Simulation of Philosophical Reflection on Hybrid Table Reasoning

Yiran Rex Ma

While LLMs demonstrate remarkable reasoning capabilities and multi-agent applicability, their tendency to "overthink" and "groupthink" pose intriguing parallels to human cognitive limitations. Inspired by this observation, we conduct an exploratory simulation to investigate whether LLMs are wise enough to be thinkers of philosophical reflection. We design two frameworks, Philosopher and Symposium, which simulate self- and group-reflection for multi-persona in hybrid table reasoning tasks. Through experiments across four benchmarks, we discover that while introducing varied perspectives might help, LLMs tend to under-perform simpler end-to-end approaches. We reveal from close reading five emergent behaviors which strikingly resemble human cognitive closure-seeking behaviors, and identify a consistent pattern of "overthinking threshold" across all tasks, where collaborative reasoning often reaches a critical point of diminishing returns. This study sheds light on a fundamental challenge shared by both human and machine intelligence: the delicate balance between deliberation and decisiveness.

Evaluating Structured Output Robustness of Small Language Models for Open Attribute-Value Extraction from Clinical Notes

Nikita Neveditsin, Pawan Lingras, Vijay Kumar Mago

We evaluate the parseability of structured outputs generated by small language models for open attribute-value extraction from clinical notes. Across JSON, YAML, and XML, JSON consistently yields the highest parseability. Targeted prompts and larger models improve structural validity, while longer documents reduce it. Error analysis reveals common failure patterns. Our findings highlight how model size, prompt specificity, document type, and output format interact to affect structural reliability.

Pun2Pun: Benchmarking LLMs on Textual-Visual Chinese-English Pun Translation via Pragmatics Model and Linguistic Reasoning

Yiran Rex Ma, Shan Huang, Yuting Xu, Ziyu Zhou, Yuanxi Wei

Puns, as a unique form of linguistic creativity, present significant challenges in cross-lingual translation, particularly between linguistically distant languages like Chinese and English, where it's often considered a "mission impossible". We introduce Pun2Pun, a novel benchmark for quantitatively evaluating pun translation between Chinese and English while preserving both linguistic mechanisms and humorous effects. We propose the adaptation of Constant-Variable Optimization (CVO) Model for translation strategy and concomitant Overlap (Ovl) metric for translation quality assessment. Our approach provides a robust quantitative evaluation framework to assess models' complex linguistic and cultural reasoning capabilities in pun translation. Through extensive experiments on both textual and visual puns, we demonstrate that our translation strategy model significantly improves performance, particularly for better-performing models. Our findings reveal exciting potentials and current limitations of LLMs in preserving sophisticated humor across linguistic and cultural boundaries.

Insights into Alignment: Evaluating DPO and its Variants Across Multiple Tasks

Amir Saeidi, Shivanshu Verma, Md Nayem Uddin, Chitta Baral

This study evaluates Direct Preference Optimization (DPO) and its variants for aligning Large Language Models (LLMs) with human preferences, testing three configurations: (1) with Supervised Fine-Tuning (SFT), (2) without SFT, and (3) without SFT but using an instruction-tuned model. We further investigate how training set size influences model performance. Our evaluation spans 13 benchmarks—covering dialogue, reasoning, mathematical problem-solving, question answering, truthfulness, MT-Bench, Big Bench, and the Open LLM Leaderboard. We find that: (1) alignment methods often achieve near-optimal performance even with smaller subsets of training data; (2) although they offer limited improvements on complex reasoning tasks, they enhance mathematical problem-solving; and (3) using an instruction-tuned model improves truthfulness. These insights highlight the conditions under which alignment methods excel, as well as their limitations.

SequentialBreak: Large Language Models Can be Fooled by Embedding Jailbreak Prompts into Sequential Prompt Chains

Bijoy Ahmed Saiem, MD Sadik Hossain Shanto, Rakib Ahsan, Md Rafi Ur Rashid

As the use of Large Language Models (LLMs) expands, so do concerns about their vulnerability to jailbreak attacks. We introduce SequentialBreak, a novel single-query jailbreak technique that arranges multiple benign prompts in sequence with a hidden malicious instruction among them to bypass safety mechanisms. Sequential prompt chains in a single query can lead LLMs to focus on certain prompts while ignoring others. By embedding a malicious prompt within a prompt chain, we show that LLMs tend to ignore the harmful context and respond to all prompts including the harmful one. We demonstrate the effectiveness of our attack across diverse scenarios—including Q\\&A systems, dialogue completion tasks, and levelwise gaming scenario—highlighting its adaptability to varied prompt structures. The variability of prompt structures shows that SequentialBreak is adaptable to formats beyond those discussed here. Experiments show that SequentialBreak only uses a single query to significantly outperform existing baselines on both open-source and closed-source models. These findings underline the urgent need for more robust defenses against prompt-based attacks. The Results and website are available on https://anonymous.4open.science/r/JailBreakAttack-4F3B/.

A Dual-Layered Evaluation of Geopolitical and Cultural Bias in LLMs

Sean Kim, Hyuhng Joon Kim

As large language models (LLMs) are increasingly deployed across diverse linguistic and cultural contexts, understanding their behavior in both factual and disputable scenarios is essential—especially when their outputs may shape public opinion or reinforce dominant narratives. This study investigates two types of bias in LLMs: model bias (bias stemming from model training) and inference bias (bias induced by the language of the query), through a two-phase evaluation. The first phase evaluates LLMs on factual questions where a single verifiable answer exists, assessing whether models maintain consistency across different query languages. The second phase expands the scope by probing geopolitically sensitive disputes, where responses may reflect culturally embedded or ideologically aligned perspectives. We construct a manually curated dataset spanning both factual and disputable QA, across four languages and question types. Our findings show that while Phase 1 reveals primarily language-induced alignment, Phase 2 reveals entangled effects of both language and model training. This study offers a structured framework for evaluating LLM behavior across neutral and sensitive topics, providing insights for future LLM deployment and culturally-aware evaluation practices in multilingual contexts.

Evaluating Credibility and Political Bias in LLMs for News Outlets in Bangladesh

Tabia Tanzin Prama, Md. Saiful Islam

Search engines increasingly use large language models (LLMs) to generate direct answers, while AI chatbots retrieve updated information from the Internet. As information curators for billions of users, LLMs must evaluate the accuracy and reliability of sources. This study audits nine LLMs from OpenAI, Google, and Meta to assess their ability to evaluate the credibility and quality of the top 20 most popular Bangladeshi news outlets. While LLMs rate most tested outlets, larger models more often refuse to rate sources due to insufficient information, while smaller models are prone to hallucinations. When ratings are provided, LLMs show strong internal consistency with an average correlation coefficient () of 0.72, but their alignment with human expert evaluations is moderate, with an average of 0.45. We introduce a dataset of expert opinions (journalism and media studies students) on the credibility and political bias of Bangladeshi news outlets to evaluate LLMs' political bias and credibility assessments. Our analysis reveals that LLMs in default configurations favor the Bangladesh Awami League-affiliated sources in credibility ratings. Assigning partisan identities to LLMs further amplifies politically congruent biases in their assessments. These findings highlight the need to address political bias and improve credibility evaluations as LLMs increasingly shape how news and political information are curated worldwide.

CHENGYU-BENCH: Benchmarking Large Language Models for Chinese Idiom Understanding and Use

Yicheng Fu, Zhemin Huang, Liuxin Yang, Yumeng Lu, Zhongdongming Dai

Chinese idioms (, Chengyu) are concise four-character expressions steeped in history and culture, whose literal translations often fail to capture their full meaning. This complexity makes them challenging for language models to interpret and use correctly. Existing benchmarks focus on narrow tasks—multiple-choice cloze tests, isolated translation, or simple paraphrasing. We introduce CHENGYU-BENCH, a comprehensive benchmark featuring three tasks: (1) Evaluative Connotation, classifying idioms as positive or negative; (2) Appropriateness, detecting incorrect idiom usage in context; and (3) Open Cloze, filling blanks in longer passages without options. CHENGYU-BENCH comprises 2,937 human-verified examples covering 1,765 common idioms sourced from diverse corpora. We evaluate leading LLMs and find they achieve over 95\% accuracy on Evaluative Connotation, but only 85% on Appropriateness and 40% top-1 accuracy in Open Cloze. Error analysis reveals that most mistakes arise from fundamental misunderstandings of idiom meanings. CHENGYU-BENCH demonstrates that while LLMs can reliably gauge idiom sentiment, they still struggle to grasp the cultural and contextual nuances essential for proper usage. The benchmark and code will be released upon paper acceptance.

Are LLMs Truly Graph-Savvy? A Comprehensive Evaluation of Graph Generation

EGE DEMRC, Rithwik Kerur, Ambuj Singh

While large language models (LLMs) have demonstrated impressive capabilities across diverse tasks, their ability to generate valid graph structures remains underexplored. We evaluate fifteen state-of-the-art LLMs on five specialized graph generation tasks spanning delivery networks, social networks, quantum circuits, gene-disease networks, and transportation systems. We also test the LLMs using 3 different prompt types: direct, iterative feedback, and program-augmented. Models supported with explicit reasoning modules (o3-mini-high, o1, Claude 3.7 Sonnet, DeepSeek-R1) solve more than twice as many tasks as their general-purpose peers, independent of parameter count. Error forensics reveals two recurring failure modes: Llama-family models often violate basic structural constraints, whereas Claude models respect topology but mismanage higher-order logical rules. Allowing models to iteratively refine their answers yields uneven gains, underscoring fundamental differences in error-correction capacity. This work demonstrates that graph competence stems from specialized architectural design rather than scale, establishing a framework for developing truly graph-savvy language models. Results and verification scripts available at https://github.com/anonymized-for-the-blind-review.

Bridging the Embodiment Gap in Agricultural Knowledge Representation for Language Models

Huijin Ju, Vasu Jindal, Zili Lyu

This paper quantifies the "embodiment gap" between disembodied language models and embodied agricultural knowledge communication through mixed-methods analysis with 78 farmers. Our key contributions include: (1) the Embodied Knowledge Representation Framework (EKRF), a novel computational architecture with specialized lexical mapping that incorporates embodied linguistic patterns from five identified domains of agricultural expertise; (2) the Embodied Prompt Engineering Protocol (EPEP), which reduced the embodiment gap by 47.3\% through systematic linguistic scaffolding techniques; and (3) the Embodied Knowledge Representation Index (EKRI), a new metric for evaluating embodied knowledge representation in language models. Implementation results show substantial improvements across agricultural domains, with particularly strong gains in tool usage discourse (58.7\%) and soil assessment terminology (67\% reduction in embodiment gap). This research advances both theoretical understanding of embodied cognition in AI and practical methodologies to enhance LLM performance in domains requiring embodied expertise.

A Framework for Fine-Grained Complexity Control in Health Answer Generation

Daniel Jorge Bernardo Ferreira, Tiago Almeida, Srgio Matos

Effective communication of health information requires adapting complexity to match the target audience's literacy level. However, manually simplifying medical content is both time-consuming and difficult to scale. To address this challenge, we developed a new framework for automatically generating health answers at multiple complexity levels. We began by collecting 166 linguistic features to quantify text complexity, including traditional readability metrics (e.g., Flesch-Kincaid, SMOG), medical terminology usage (e.g., UMLS coverage, medical entity recognition), syntactic complexity, semantic coherence, and LLM-based measures (e.g., masked language modeling, LLM-as-a-judge). Applying these features to a custom dataset of parallel health texts and external medical benchmarks, we used feature selection to identify 13 key metrics that reliably distinguish between simple and complex text pairs. From these, we derived a complexity scoring formula by combining the metrics with weights learned from a logistic regression model. Using this formula, we created a large multi-level dataset of health question-answer pairs, ranging from elementary patient-friendly explanations to advanced technical summaries. The initial QA pairs came from established datasets including LiveQA, MedicationQA, and MEDIQA-AnS. We then used LLaMA-based language models with carefully engineered prompts to transform the original answers into five different versions ordered by complexity. Finally, we fine-tuned a large language model on this dataset, incorporating special tokens to control the complexity of the generated text. The resulting model can generate health answers at fine-grained complexity levels, allowing users to select the desired level of detail and technicality.

Time-LlaMA: Adapting Large Language Models for Time Series Modeling via Dynamic Low-rank Adaptation

Ellen Yi Ge, Juyuan Zhang, Wei Zhu, Hui Yi Leong, Jiechao Gao

Time series modeling holds significant importance in many industrial applications and has been extensively studied. A series of recent studies have demonstrated that large language models (LLMs) possess robust pattern recognition and semantic understanding capabilities over time series data. However, the current literature have yet striked a high-quality balance between (a) effectively aligning the time series and natural language modalities and (b) keeping the inference efficiency for industrial deployment. To address the above issues, we now propose the Time-LlaMA framework. Time-LlaMA first converts the time series input into token embeddings through a linear tokenization mechanism. Second, the time series token embeddings are aligned with the text prompts. Third, to further adapt the large languag model (LLM) backbone for time series modeling, we have developed a dynamic low-rank adaptation technique (DynaLoRA). DynaLoRA dynamically chooses the most suitable LoRA modules at each layer of the Transformer backbone for each time series input, enhancing the model's predictive capabilities. Our experimental results on an extensive collection of challenging open and proprietary time series tasks confirm that our proposed method achieves the state-of-the-art (SOTA) performance and have potentials for wide industrial usages.\footnote{Codes will be made public upon acceptance. }

RusConText Benchmark: A Russian Language Evaluation Benchmark for Understanding Context

Andrey Chirkin, Svetlana Kuznetsova, Maria Volina, Anna V. Dengina

This paper represents an implementation of an approach rather similar to that of (Zhu et al., 2024), but for the Russian-language data. We introduce a RusConText Benchmark for evaluating short context understanding in Russian, comprising four distinct yet interrelated tasks: ellipsis resolution, coreference resolution, and idiom interpretation, and discourse understanding. Each task targets a specific aspect of linguistic processing, challenging large language model to recover omitted information, resolve referential dependencies, interpret idioms and discourse. RusConText Benchmark serves as an additional resource beyond standard benchmarks, designed to assess model performance from a specific perspective. In addition, we present the results of scoring 4 models on our benchmark.

Supervised Neural Topic Modeling with Label Alignment

Ruihao Chen, Hegang Chen, Yuyin Lu, Yanghui Rao, Chunjiang Zhu

Neural topic modeling is a scalable automated technique for text data mining. In various downstream tasks of topic modeling, it is prefered that the discovered topics well align with labels. However, due to the lack of guidance from labels, unsupervised neural topic models are less powerful in this situation. Existing supervised neural topic models often adopt a label-free prior to generate the latent document-topic distributions and use them to predict the labels and thus achieve label-topic alignment indirectly. Such a mechanism faces the following issues: 1) The label-free prior leads to topics blending the latent patterns of multiple labels; 2) One is unable to intuitively identify the explicit relationships between labels and the discovered topics. To tackle these problems, we develop a novel supervised neural topic model which utilizes a chain-structured graphical model that emphasizes direct constraints of labels on topics. Soft indicators are introduced to explicitly construct the label-topic relationships. To obtain well-organized label-topic relationships, we formalize an entropy-regularized optimal transport problem on the embedding space and model them as the transport plan. Moreover, our proposed method can be seamlessly integrated with most existing unsupervised neural topic models in a flexible manner. Experiment results on multiple datasets demonstrate that our model can greatly enhance the alignment between labels and topics while maintaining good topic quality.

Patchwise Cooperative Game-based Interpretability Method for Large Vision-Language Models

Yao Zhu, Yunjian Zhang, Zizhe Wang, Xiu Yan, Peng Sun, Xiangyang Ji
Amidst the rapid advancement of artificial intelligence, research on large vision-language models (LVLMs) has emerged as a pivotal area. However, understanding their internal mechanisms remains challenging due to the limitations of existing interpretability methods, especially regarding faithfulness and plausibility. To address this, we first construct a human response interpretability dataset that evaluates the plausibility of model explanations by comparing the attention regions between the model and humans when answering the same questions. We then propose a patchwise cooperative game-based interpretability method for LVLMs, which employs Shapley values to quantify the impact of individual image patches on generation likelihood and enhances computational efficiency through a single input approximation approach. Experimental results demonstrate our method's faithfulness, plausibility, and robustness. Our method provides researchers with deeper insights into model behavior, allowing for an examination of the specific image regions each layer relies on during response generation, ultimately enhancing model reliability. Our codes are available at https://github.com/ZY123-GOOD/Patchwise_Cooperative.

Data Contamination Quiz: A Tool to Detect and Estimate Contamination in Large Language Models

Shahriar Golchin, Mihai Surdeanu
We propose the Data Contamination Quiz (DCQ), a simple and effective approach to detect data contamination in large language models (LLMs) and estimate the amount of it. Specifically, we frame data contamination detection as a series of multiple-choice questions, devising a quiz format wherein three perturbed versions of each instance, subsampled from a specific dataset partition, are created. These changes only include word-level perturbations. The generated perturbations, along with the original dataset instance, form the options in the DCQ, with an extra option accommodating the selection of none of the provided options. Given that the only distinguishing signal among the options is the exact wording with respect to the original dataset instance, an LLM, when tasked with identifying the original dataset instance, gravitates towards selecting the original one if it has been exposed to it. While accounting for positional biases in LLMs, the quiz performance reveals the contamination level for the tested model with the dataset partition to which the quiz pertains. Applied to various datasets and LLMs, under controlled and uncontrolled contamination, our findings—while fully lacking access to training data and model parameters—suggest that DCQ achieves state-of-the-art results and uncovers greater contamination/memorization levels compared to existing methods. Also, it proficiently bypasses more safety filters, especially those set to avoid generating copyrighted content.

A Parallelized Framework for Simulating Large-Scale LLM Agents with Realistic Environments and Interactions

Jun Zhang, Yuwei Yan, Junbo Yan, Zhiheng Zheng, Jinghua Piao, Depeng Jin, Yong Li
The development of large language models (LLMs) offers a feasible approach to simulating complex behavioral patterns of individuals, enabling the reconstruction of microscopic and realistic human societal dynamics. However, this approach demands a realistic environment to provide feedback for the evolving of agents, as well as a parallelized framework to support the massive and uncertain interactions among agents and environments. To address the gaps in existing works, which lack real-world environments and struggle with complex interactions, we design a scalable framework named **AgentSociety**, which integrates realistic societal environments and parallelized interactions to support simulations of large-scale agents. Experiments demonstrate that the framework can support simulations of 30,000 agents that are faster than the wall-clock time with 24 NVIDIA A800 GPUs and the performance grows linearly with the increase of LLM computational resources. We also show that the integration of realistic environments significantly enhances the authenticity of the agents' behaviors. Through the framework and experimental results, we are confident that deploying large-scale LLM Agents to simulate human societies becomes feasible. This will help practitioners in fields such as social sciences and management sciences to obtain new scientific discoveries via language generation technologies, and even improve planning and decision-making in the real world. The code is available at https://github.com/tsinghua-fib-lab/agentsociety/.

ENGinius: A Bilingual LLM Optimized for Plant Construction Engineering

Wooseong Lee, Minseo Kim, Taeil Hur, Gyeong Hwan Jang, Woncheol Lee, Maro Na, Taeuk Kim

Recent advances in large language models (LLMs) have drawn attention for their potential to automate and optimize processes across various sectors. However, the adoption of LLMs in the plant construction industry remains limited, mainly due to its highly specialized nature and the lack of resources for domain-specific training and evaluation. In this work, we propose ENGinius, the first LLM designed for plant construction engineering. We present procedures for data construction and model training, along with the first benchmarks tailored to this underrepresented domain. We show that ENGinius delivers optimized responses to plant engineers by leveraging enriched domain knowledge. We also demonstrate its practical impact and use cases, such as technical document processing and multilingual communication.

TaDA: Training-free recipe for Decoding with Adaptive KV Cache Compression and Mean-centering

Vinay Joshi, Pratik Prabhanjan Brahma, Zicheng Liu, Emad Barsoum

The key-value (KV) cache in transformer models is a critical component for efficient decoding or inference, yet its memory demands scale poorly with sequence length, posing a major challenge for scalable deployment of large language models. Among several approaches to KV cache compression, quantization of key and value activations has been widely explored. Most KV cache quantization methods still need to manage sparse and noncontiguous outliers separately. To address this, we introduce TaDA, a training-free recipe for KV cache compression with quantization precision that adapts to error sensitivity across layers and a mean centering to eliminate separate outlier handling. Our approach yields substantial accuracy improvements for multiple models supporting various context lengths. Moreover, our approach does not need to separately manage outlier elements—a persistent hurdle in most traditional quantization methods. Experiments on standard benchmarks demonstrate that our technique reduces KV cache memory footprint to 27% of the original 16-bit baseline while achieving comparable accuracy. Our method paves the way for scalable and high-performance reasoning in language models by potentially enabling inference for longer context length models, reasoning models, and longer chain of thoughts.

EXPLAIN: Enhancing Retrieval-Augmented Generation with Entity Summary

Yaozhen Liang, Xiao Liu, Jiajun Yu, Fang Zhouhua, Qunsheng Zou, Linghan Zheng, Yong Li, Zhiwei Liu, Haishuai Wang

Document question answering plays a crucial role in enhancing employee productivity by providing quick and accurate access to information. Two primary approaches have been developed: retrieval-augmented generation (RAG), which reduces input tokens and inference costs, and long-context question answering (LC), which processes entire documents for higher accuracy. We introduce EXPLAIN (\textbf{EX}tracting, \textbf{P}re-summarizing, \textbf{L}inking and enh\textbf{A}c\textbf{IN}g RAG), a novel retrieval-augmented generation method that automatically extracts useful entities and generates summaries from documents. EXPLAIN improves accuracy by retrieving more informative entity summaries, achieving precision comparable to LC while maintaining low token consumption. Experimental results on internal dataset (ROUGE-L from 30.14\% to 30.31\%) and three public datasets (HotpotQA, 2WikiMQA, and Quality, average score from 62\% to 64\%) demonstrate the efficacy of EXPLAIN. Human evaluation in ant group production deployment indicates EXPLAIN surpasses baseline RAG in comprehensiveness.

Capture the Key in Reasoning to Enhance CoT Distillation Generalization

Chengwei Dai, Kun Li, Wei Zhou, Songlin Hu

As Large Language Models (LLMs) scale up and gain powerful Chain-of-Thoughts (CoTs) reasoning abilities, practical resource constraints drive efforts to distill these capabilities into more compact Smaller Language Models (SLMs). We find that CoTs consist mainly of simple reasoning forms, with a small proportion (\4.7%) of key reasoning steps that truly impact conclusions. However, previous distillation methods typically involve supervised fine-tuning student SLMs only on correct CoTs data produced by teacher LLMs, resulting in students struggling to learn the key, instead imitating the teacher's reasoning forms and making errors or omissions in reasoning. To address these issues, drawing an analogy to human learning, where analyzing mistakes according to correct solutions often reveals the crucial steps leading to successes or failures, we propose mistak\textbf{E}-\textbf{D}riven key reason\textbf{I}ng step distilla\textbf{T}ion (\textbf{EDIT}), a novel method that further aids SLMs learning key reasoning steps rather than mere simple fine-tuning. Firstly, to expose the crucial steps in CoTs, we carefully design specific prompts to generate dual CoTs data with similar reasoning paths but divergent conclusions. Then, we apply the minimum edit distance algorithm on the dual CoTs data to locate these key steps and optimize the likelihood on these tokens. Extensive experiments and analysis validate the effectiveness of EDIT across both in-domain(IND) and out-of-domain(OOD) benchmark reasoning datasets\footnote{Code can be found at }.

Direct Prompt Optimization with Continuous Representations

Yangkun Wang, Zihan Wang, Jingbo Shang

Prompt optimization for language models faces challenges due to the large discrete search space, the reliance on continuous gradient updates, and the need to round continuous representations into discrete prompts, which causes inflexibility and instability. Existing methods attempt to address these by constraining the search space and adopting greedy, incremental improvements, but they often fail to fully leverage historical gradient information. In this paper, we model the prompt optimization problem by the probability distribution of the prompt and present a novel approach that integrates greedy strategies into optimization with continuous representations. This approach can exploit historical gradient information to address the instability caused by rounding in existing methods. Our study indicates that using continuous representations can improve prompt optimization performance on both text classification and attack tasks, as well as models, including GPT-2, OPT, Vicuna, and LLaMA-2, and also be adaptable to models of different sizes.

An Efficient Task-Oriented Dialogue Policy: Evolutionary Reinforcement Learning Injected by Elite Individuals.

Yangyang Zhao, Ben Niu, Libo Qin, Shihan Wang

Deep Reinforcement Learning (DRL) is widely used in task-oriented dialogue systems to optimize dialogue policy, but it struggles to balance exploration and exploitation due to the high dimensionality of state and action spaces. This challenge often results in local optima or poor convergence. Evolutionary Algorithms (EAs) have been proven to effectively explore the solution space of neural networks by maintaining population diversity. Inspired by this, we innovatively combine the global search capabilities of EA with the local optimization of DRL to achieve a balance between exploration and exploitation. Nevertheless, the inherent flexibility of natural language in dialogue tasks complicates this direct integration, leading to prolonged evolutionary times. Thus, we further propose an elite individual injection mechanism to enhance EA's search efficiency by adaptively introducing best-performing individuals into the population. Experiments across four datasets show that our approach significantly improves the balance between exploration and exploitation, boosting performance. Moreover, the effectiveness of the EII mechanism in reducing exploration time has been demonstrated, achieving an efficient integration of EA and DRL on task-oriented dialogue policy tasks.

Sharper and Faster mean Better: Towards More Efficient Vision-Language Model for Hour-scale Long Video Understanding

Daoze Zhang, Yuze Zhao, Jintao Huang, Yingda Chen

Despite existing multimodal language models showing impressive performance on the video understanding task, extremely long videos still pose significant challenges to language model's context length, memory consumption, and computational complexity. To address these issues, we propose a vision-language model named Sophia for long video understanding, which can efficiently handle hour-scale long videos. First, we employ a Shot-adaptive Frame Pruning technique, which naturally segments long videos into multiple camera shots, to more sharply identify and focus on the frames relevant to the query. Additionally, we introduce a Hierarchical Attention mechanism to effectively model the long-term temporal dependencies between video frames, which achieves a time and space complexity of O(N) w.r.t. the input sequence length N while theoretically maintaining the global modeling efficiency. Experimentally, our Sophia exhibits competitive performance compared to existing video understanding baselines across various benchmarks for long video understanding with reduced time and memory consumption. The model code and weights are available at https://huggingface.co/Tao-tse/Sophia.

GainRAG: Preference Alignment in Retrieval-Augmented Generation through Gain Signal Synthesis

Yi Jiang, Sendong Zhao, Jianbo Li, Haochun Wang, Bing Qin

The Retrieval-Augmented Generation (RAG) framework introduces a retrieval module to dynamicaslly inject retrieved information into the input context of large language models (LLMs), and has demonstrated significant success in various NLP tasks. However, the current study points out that there is a preference gap between retrievers and LLMs in the RAG framework, which limit the further improvement of system performance. Some highly relevant passages may interfere with LLM reasoning because they contain complex or contradictory information; while some indirectly related or even inaccurate content may help LLM generate more accurate answers by providing suggestive information or logical clues. To solve this, we propose **GainRAG**, a novel approach that aligns the retriever’s and LLM’s preferences by defining a new metric, “gain’’, which measure how well an input passage contributes to correct outputs. We then propose a method to estimate these gain signals and train a middleware that aligns the preferences of the retriever and the LLM using only limited data. In addition, we introduce a pseudo-passage strategy to mitigate degradation. The experimental results on 6 datasets verify the effectiveness of GainRAG.

Less Mature is More Adaptable for Sentence-level Language Modeling

Abhilasha Sancheti, David Dale, Artyom Kozhevnikov, Maha Elbayad

This work investigates sentence-level models (\textit{i.e.}, models that operate at the sentence-level) to study how sentence representations from various encoders influence downstream task performance, and which syntactic, semantic, and discourse-level properties are essential for strong performance. Our experiments encompass encoders with diverse training regimes and pretraining domains, as well as various pooling strategies applied to multi-sentence input tasks (including sentence ordering, sentiment classification, and natural language inference) requiring coarse-to-fine-grained reasoning. We find that "less mature" representations (\textit{e.g.}, mean-pooled representations from BERT's first or last layer, or representations from encoders with limited fine-tuning) exhibit greater generalizability and adaptability to downstream tasks compared to representations from extensively fine-tuned models (\textit{e.g.}, SBERT or SimCSE). These findings are consistent across different pretraining seed initializations for BERT. Our probing analysis reveals that syntactic and discourse-level properties are stronger indicators of downstream performance than MTEB scores or decodability. Furthermore, the data and time efficiency of sentence-level models, often outperforming token-level models, underscores their potential for future research.

What are the Essential Factors in Crafting Effective Long Context Multi-Hop Instruction Datasets? Insights and Best Practices

Zhi Chen, Qiguang Chen, Libo Qin, Qipeng Guo, haijun Lv, Yicheng Zou, Hang Yan, Kai Chen, Dahua Lin

Recent advancements in large language models (LLMs) with extended context windows have significantly improved various tasks. To improve long-context capabilities, much work focuses on augmenting LLM's capabilities with synthetic data. Existing methods often leverage the Self-Instruct framework to generate long-context instruction-tuning data. However, our preliminary experiments show that fewer than 35% of samples generated by Qwen-2-72B are multi-hop, and over 40% exhibit poor quality, limiting comprehensive understanding and further research. To address this, we propose the Multi-agent Interactive Multi-hop Generation (MIMG) framework, which integrates a quality verification agent, a single-hop question generation agent, a multiple question sampling strategy, and a multi-hop question merger agent. This framework significantly improves data quality, with high-quality, multi-hop, and diverse data. Furthermore, we conduct a thorough analysis of document selection, question merging, and validation techniques through extensive experiments across various models. Our results demonstrate that synthetic high-quality long-context instruction data can enhance model performance, surpassing even models trained on larger amounts of human-annotated data.

Speed Up Your Code: Progressive Code Acceleration Through Bidirectional Tree Editing

Longhui Zhang, Jiahao Wang, Meishan Zhang, GaoXiong Cao, Ensheng Shi, mayuchi, Jun Yu, Honghai LIU, Jing Li, Min Zhang

Large language models (LLMs) have made significant strides in code acceleration (CA) tasks. Current works typically fine-tune LLMs using slow-fast code pairs mined from online programming platforms. Although these methods are widely recognized for their effectiveness, the training data often lack clear code acceleration patterns and offer only limited speed improvements. Moreover, existing training methods, such as direct instruction fine-tuning (IFT), tend to overlook the hierarchical relationships among acceleration patterns. In this work, we introduce BITE, a novel training paradigm designed to improve LLMs' CA capabilities through two key innovations: (1) Bidirectional tree editing, which generates high-quality training data by incrementally transforming given code into both its most efficient and least efficient variants, and (2) Progressive code acceleration learning, which enables LLMs to internalize multi-level CA strategies by learning increasingly sophisticated acceleration patterns. Additionally, we introduce a new CA evaluation benchmark and metric for comprehensive assessment of model performance on CA tasks. Extensive experiments on both our benchmark and existing benchmarks demonstrate the effectiveness of our approach. Notably, BITE enables Qwen-1.5B to outperform prompt-enhanced GPT-4 and current training-based methods on average across five programming languages.

Rethinking Semantic Parsing for Large Language Models: Enhancing LLM Performance with Semantic Hints

Kaikai An, Shuzheng Si, Helan Hu, Haozhe Zhao, Yuchi Wang, Qingyan Guo, Baobao Chang

Semantic Parsing aims to capture the meaning of a sentence and convert it into a logical, structured form. Previous studies show that semantic parsing enhances the performance of smaller models (e.g., BERT) on downstream tasks. However, it remains unclear whether the improvements extend similarly to LLMs. In this paper, our empirical findings reveal that, unlike smaller models, directly adding semantic parsing results into LLMs reduces their performance. To overcome this, we propose SENSE, a novel prompting approach that embeds semantic hints within the prompt. Experiments show that SENSE consistently improves LLMs' performance across various tasks, highlighting the potential of integrating semantic information to improve LLM capabilities.

Bilingual Zero-Shot Stance Detection

Chenye Zhao, Cornelia Caragea

Zero-shot stance detection (ZSSD) aims to determine whether the author of a text is in support, against, or neutral toward a target that is unseen during training. In this paper, we investigate ZSSD within a bilingual framework and compare it with cross-lingual and monolingual scenarios, in settings that have not previously been explored. Our study focuses on both noun-phrase and claim targets within in-domain and out-of-domain bilingual ZSSD scenarios. To support this research, we assemble Bi-STANCE, a comprehensive bilingual ZSSD dataset consisting of over 100,000 annotated text-target pairs in both Chinese and English, sourced from existing datasets. Additionally, we examine a more challenging aspect of bilingual ZSSD by focusing on claim targets with a low occurrence of shared words with their corresponding texts. As part of Bi-STANCE, we created an extended dataset that emphasizes this challenging scenario. To the best of our knowledge, we are the first to explore this difficult ZSSD setting. We investigate these tasks using state-of-the-art pre-trained language models (PLMs) and large language models (LLMs). We release our dataset and code at https://github.com/chenyez/BiSTANCE.

CCHall: A Novel Benchmark for Joint Cross-Lingual and Cross-Modal Hallucinations Detection in Large Language Models

Yongheng Zhang, Xu Liu, Ruoxi Zhou, Qiguang Chen, Hao Fei, Wenpeng Lu, Libo Qin

Investigating hallucination issues in large language models (LLMs) within cross-lingual and cross-modal scenarios can greatly advance the large-scale deployment in real-world applications. Nevertheless, the current studies are limited to a single scenario, either cross-lingual or cross-modal, leaving a gap in the exploration of hallucinations in the joint cross-lingual and cross-modal scenarios. Motivated by this, we introduce a novel joint Cross-lingual and Cross-modal Hallucinations benchmark (CCHall) to fill this gap. Specifically, CCHall simultaneously incorporates both cross-lingual and cross-modal hallucination scenarios, which can be used to assess the cross-lingual and cross-modal capabilities of LLMs. Furthermore, we conduct a comprehensive evaluation on CCHall, exploring both mainstream open-source and closed-source LLMs. The experimental results highlight that current LLMs still struggle with CCHall. We hope CCHall can serve as a valuable resource to assess LLMs in joint cross-lingual and cross-modal scenarios.

11

Day 3: Wednesday 30 July

Oral Presentation

Session 11: Computational Social Science and Cultural Analytics (Wednesday 30 July, 09:00-10:30, Hall C)

09:11 - 09:22

The AI Gap: How Socioeconomic Status Affects Language Technology Interactions

Elisa Bassignana, Amanda Cercas Curry, Dirk Hovy

Elisa Bassignana, Amanda Cercas Curry, Dirk Hovy

09:44 - 09:55

Culture is Not Trivia: Sociocultural Theory for Cultural NLP

Naitian Zhou, David Bamman, Isaac L. Bleaman

Naitian Zhou, David Bamman, Isaac L. Bleaman

Session 11: Efficient/Low-Resource Methods for NLP (Wednesday 30 July, 09:00-10:30, Hall A)

09:44 - 09:55

SubLIME: Subset Selection via Rank Correlation Prediction for Data-Efficient LLM Evaluation

Gayathri Saranathan, Cong Xu, Mahammad Parwez Alam, Tarun Kumar, Martin Foltin, Soon Yee Wong, Suparna Bhattacharya

Gayathri Saranathan, Cong Xu, Mahammad Parwez Alam, Tarun Kumar, Martin Foltin, Soon Yee Wong, Suparna Bhattacharya

Session 11: Efficient/Low-Resource Methods for NLP (Wednesday 30 July, 09:00-10:30, Room 1.62)

09:00 - 09:11

FedEx-LoRA: Exact Aggregation for Federated and Efficient Fine-Tuning of Large Language Models

Raghav Singhal, Kaustubh Ponkshe, Praneeth Vepakomma

Raghav Singhal, Kaustubh Ponkshe, Praneeth Vepakomma

09:11 - 09:22

Turning Trash into Treasure: Accelerating Inference of Large Language Models with Token Recycling

Xianzhen Luo, Yixuan Wang, Qingfu Zhu, Zhiming Zhang, Xuanyu Zhang, Qing Yang, Dongliang Xu

Xianzhen Luo, Yixuan Wang, Qingfu Zhu, Zhiming Zhang, Xuanyu Zhang, Qing Yang, Dongliang Xu

09:22 - 09:33

SCOPE: Optimizing Key-Value Cache Compression in Long-context Generation

Jialong Wu, Zhenglin Wang, Linhai Zhang, Yilong Lai, Yulan He, Deyu Zhou

Jialong Wu, Zhenglin Wang, Linhai Zhang, Yilong Lai, Yulan He, Deyu Zhou

09:33 - 09:44

Meta-rater: A Multi-dimensional Data Selection Method for Pre-training Language Models

Xinlin Zhuang, Jiahui Peng, Ren Ma, Yinfan Wang, Tianyi Bai, Xingjian Wei, Qiu Jiantao, Chi Zhang, Ying Qian, Conghui He

Xinlin Zhuang, Jiahui Peng, Ren Ma, Yinfan Wang, Tianyi Bai, Xingjian Wei, Qiu Jiantao, Chi Zhang, Ying Qian, Conghui He

09:44 - 09:55

Adapt Once, Thrive with Updates: Transferable Parameter-Efficient Fine-Tuning on Evolving Base Models

Naibin Gu, Peng Fu, Xiyu Liu, Ke Ma, Zheng Lin, Weiping Wang

Naibin Gu, Peng Fu, Xiyu Liu, Ke Ma, Zheng Lin, Weiping Wang

09:55 - 10:06

Towards the Law of Capacity Gap in Distilling Language Models

Chen Zhang, Qiuchi Li, Dawei Song, Zheyu Ye, Yan Gao, Yao Hu

Chen Zhang, Qiuchi Li, Dawei Song, Zheyu Ye, Yan Gao, Yao Hu

10:06 - 10:17

One QuantLLM for ALL: Fine-tuning Quantized LLMs Once for Efficient Deployments

Ke Yi, Yuhui Xu, Heng Chang, Yuan Meng, Tong Zhang, Jia Li

Ke Yi, Yuhui Xu, Heng Chang, Yuan Meng, Tong Zhang, Jia Li

10:17 - 10:28

"Give Me BF16 or Give Me Death"? Accuracy-Performance Trade-Offs in LLM Quantization

Eldar Kurtic, Alexandre Noll Marques, Shubhra Pandit, Mark Kurtz, Dan Alistarh

Eldar Kurtic, Alexandre Noll Marques, Shubhra Pandit, Mark Kurtz, Dan Alistarh

Session 11: Ethics, Bias, and Fairness (Wednesday 30 July, 09:00-10:30, Hall C)

09:00 - 09:11

Fairness through Difference Awareness: Measuring \textit{Desired} Group Discrimination in LLMs

Angelina Wang, Michelle Phan, Daniel E. Ho, Sanmi Koyejo

Angelina Wang, Michelle Phan, Daniel E. Ho, Sanmi Koyejo

Session 11: Human-Centered NLP (Wednesday 30 July, 09:00-10:30, Hall B)

09:44 - 09:55

Unintended Harms of Value-Aligned LLMs: Psychological and Empirical Insights

Sooyung Choi, Jaehyeok Lee, Xiaoyuan Yi, Jing Yao, Xing Xie, JinYeong Bak

Sooyung Choi, Jaehyeok Lee, Xiaoyuan Yi, Jing Yao, Xing Xie, JinYeong Bak

Session 11: Human-Centered NLP (Wednesday 30 July, 09:00-10:30, Hall C)

09:33 - 09:44

Are Rules Meant to be Broken? Understanding Multilingual Moral Reasoning as a Computational Pipeline with UniMoral

Shivani Kumar, David Jurgens

Shivani Kumar, David Jurgens

Session 11: Interpretability and Analysis of Models for NLP (Wednesday 30 July, 09:00-10:30, Hall N.1)

09:00 - 09:11

A Theory of LLM Sampling: Part Descriptive and Part Prescriptive

Sarath Sivaprasad, Pramod Kaushik, Sahar Abdelnabi, Mario Fritz

Sarath Sivaprasad, Pramod Kaushik, Sahar Abdelnabi, Mario Fritz

09:11 - 09:22

Mapping 1,000+ Language Models via the Log-Likelihood Vector

Momose Oyama, Hiroaki Yamagiwa, Yusuke Takase, Hidetoshi Shimodaira

Momose Oyama, Hiroaki Yamagiwa, Yusuke Takase, Hidetoshi Shimodaira

09:22 - 09:33

Capability Salience Vector: Fine-grained Alignment of Loss and Capabilities for Downstream Task Scaling Law

Qiming Ge, Shuhao Xing, Songyang Gao, Yunhua Zhou, Yicheng Zou, Songyang Zhang, Zhi Chen, Hang Yan, Qi Zhang, Qipeng Guo, Kai Chen

Qiming Ge, Shuhao Xing, Songyang Gao, Yunhua Zhou, Yicheng Zou, Songyang Zhang, Zhi Chen, Hang Yan, Qi Zhang, Qipeng Guo, Kai Chen

09:33 - 09:44

Llama See, Llama Do: A Mechanistic Perspective on Contextual Entrainment and Distraction in LLMs

Jingcheng Niu, Xingdi Yuan, Tong Wang, Hamidreza Saghir, Amir H. Abdi

Jingcheng Niu, Xingdi Yuan, Tong Wang, Hamidreza Saghir, Amir H. Abdi

09:44 - 09:55

Learnability on the Information-Theoretic Continuum: Inductive Bias for Information Locality in Neural Language Models

Taiga Someya, Anej Svete, Brian DuSell, Timothy J. O'Donnell, Mario Giulianelli, Ryan Cotterell

Taiga Someya, Anej Svete, Brian DuSell, Timothy J. O'Donnell, Mario Giulianelli, Ryan Cotterell

Session 11: Interpretability and Analysis of Models for NLP (Wednesday 30 July, 09:00-10:30, Room 1.85)

09:00 - 09:11

Influences on LLM Calibration: A Study of Response Agreement, Loss Functions, and Prompt Styles

Yuxi Xia, Pedro Henrique Luz de Araujo, Klim Zaporojets, Benjamin Roth

Yuxi Xia, Pedro Henrique Luz de Araujo, Klim Zaporojets, Benjamin Roth

09:11 - 09:22

Sparse Latents Steer Retrieval-Augmented Generation

Chunlei Xin, Shuheng Zhou, Huijia Zhu, Weiqiang Wang, Xuanang Chen, Xinyan Guan, Yaojie Lu, Hongyu Lin, Xianpei Han, Le Sun

Chunlei Xin, Shuheng Zhou, Huijia Zhu, Weiqiang Wang, Xuanang Chen, Xinyan Guan, Yaojie Lu, Hongyu Lin, Xianpei Han, Le Sun

09:22 - 09:33

Geometric Signatures of Compositionality Across a Language Model's Lifetime

Jin Hwa Lee, Thomas Jiralerspong, Lei Yu, Yoshua Bengio, Emily Cheng

Jin Hwa Lee, Thomas Jiralerspong, Lei Yu, Yoshua Bengio, Emily Cheng

09:33 - 09:44

Finite State Automata Inside Transformers with Chain-of-Thought: A Mechanistic Study on State Tracking
Yifan Zhang, Wenyu Du, Dongming Jin, Jie Fu, Zhi Jin
Yifan Zhang, Wenyu Du, Dongming Jin, Jie Fu, Zhi Jin

09:44 - 09:55

Why Safeguarded Ships Run Aground? Aligned Large Language Models' Safety Mechanisms Tend to Be Anchored in The Template Region
Chak Tou Leong, Qingyu Yin, Jian Wang, Wenjie Li
Chak Tou Leong, Qingyu Yin, Jian Wang, Wenjie Li

09:55 - 10:06

Inferring Functionality of Attention Heads from their Parameters
Amit Elhelo, Mor Geva
Amit Elhelo, Mor Geva

10:06 - 10:17

LaTIM: Measuring Latent Token-to-Token Interactions in Mamba Models
Hugo Pitorro, Marcos Vinicius Treviso
Hugo Pitorro, Marcos Vinicius Treviso

10:17 - 10:28

Unique Hard Attention: A Tale of Two Sides
Selim Jerad, Anej Svete, Jiaoda Li, Ryan Cotterell
Selim Jerad, Anej Svete, Jiaoda Li, Ryan Cotterell

Session 11: Language Modeling (Wednesday 30 July, 09:00-10:30, Hall B)

09:00 - 09:11

Intuitive Fine-Tuning: Towards Simplifying Alignment into a Single Process
Ermo Hua, Biqing Qi, Kaiyan Zhang, Kai Tian, Xingtai Lv, Ning Ding, Bowen Zhou
Ermo Hua, Biqing Qi, Kaiyan Zhang, Kai Tian, Xingtai Lv, Ning Ding, Bowen Zhou

09:11 - 09:22

Nudging: Inference-time Alignment of LLMs via Guided Decoding
Yu Fei, Yasaman Razeghi, Sameer Singh
Yu Fei, Yasaman Razeghi, Sameer Singh

09:33 - 09:44

Language Models Resist Alignment: Evidence From Data Compression
Jiaming Ji, Kaile Wang, Tianyi Qiu, Boyuan Chen, Jiayi Zhou, Changye Li, Hantao Lou, Josef Dai, Yunhuai Liu, Yaodong Yang
Jiaming Ji, Kaile Wang, Tianyi Qiu, Boyuan Chen, Jiayi Zhou, Changye Li, Hantao Lou, Josef Dai, Yunhuai Liu, Yaodong Yang

Session 11: Language Modeling (Wednesday 30 July, 09:00-10:30, Room 1.15-16)

09:00 - 09:11

Transforming Common Crawl into a Refined Long-Horizon Pretraining Dataset

Dan SU, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro

Dan SU, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro

09:11 - 09:22

FoldMoE: Efficient Long Sequence MoE Training via Attention-MoE Pipelining

Guichao Zhu, Lintian Lei, Yuhao QING, Yichao Fu, Fanxin Li, Dong HUANG, Zekai Sun, Heming Cui

Guichao Zhu, Lintian Lei, Yuhao QING, Yichao Fu, Fanxin Li, Dong HUANG, Zekai Sun, Heming Cui

09:22 - 09:33

YuLan-Mini: Pushing the Limits of Open Data-efficient Language Model

Hu Yiwen, Song Huatong, Jie Chen, Jia Deng, jiapeng wang, Kun Zhou, Yutao Zhu, Jinhao Jiang, zican Dong, YANG LU, Xu Miao, Xin Zhao, Ji-Rong Wen

Hu Yiwen, Song Huatong, Jie Chen, Jia Deng, jiapeng wang, Kun Zhou, Yutao Zhu, Jinhao Jiang, zican Dong, YANG LU, Xu Miao, Xin Zhao, Ji-Rong Wen

09:33 - 09:44

Attention Entropy is a Key Factor: An Analysis of Parallel Context Encoding with Full-attention-based Pre-trained Language Models

Zhisong Zhang, Yan Wang, Xinting Huang, Tianqing Fang, Hongming Zhang, Chenlong Deng, Shuaiyi Li, Dong Yu

Zhisong Zhang, Yan Wang, Xinting Huang, Tianqing Fang, Hongming Zhang, Chenlong Deng, Shuaiyi Li, Dong Yu

09:44 - 09:55

LongReD: Mitigating Short-Text Degradation of Long-Context Large Language Models via Restoration Distillation

Zican Dong, Junyi Li, Jinhao Jiang, Mingyu Xu, Xin Zhao, Bingning Wang, Weipeng Chen

Zican Dong, Junyi Li, Jinhao Jiang, Mingyu Xu, Xin Zhao, Bingning Wang, Weipeng Chen

09:55 - 10:06

Acquisition and Application of Novel Knowledge in Large Language Models

Ziyu Shang, Jianghan Liu, Zhizhao Luo, Peng Wang, Wenjun Ke, Jiajun Liu, Zijie Xu, Guozheng Li

Ziyu Shang, Jianghan Liu, Zhizhao Luo, Peng Wang, Wenjun Ke, Jiajun Liu, Zijie Xu, Guozheng Li

10:06 - 10:17

Tokenisation is NP-Complete

Philip Whittington, Gregor Bachmann, Tiago Pimentel

Philip Whittington, Gregor Bachmann, Tiago Pimentel

Session 11: Machine Learning for NLP (Wednesday 30 July, 09:00-10:30, Room 1.61)

09:00 - 09:11

Mitigating Shortcut Learning with InterpoLated Learning
Michalis Korakakis, Andreas Vlachos, Adrian Weller
Michalis Korakakis, Andreas Vlachos, Adrian Weller

09:11 - 09:22

Rethinking Reward Model Evaluation Through the Lens of Reward Overoptimization
Sunghwan Kim, Dongjin Kang, Taeyoon Kwon, Hyungjoo Chae, Dongha Lee, Jinyoung Yeo
Sunghwan Kim, Dongjin Kang, Taeyoon Kwon, Hyungjoo Chae, Dongha Lee, Jinyoung Yeo

09:22 - 09:33

Automatic Expert Discovery in LLM Upcycling via Sparse Interpolated Mixture-of-Experts
Shengzhuang Chen, Ying Wei, Jonathan Richard Schwarz
Shengzhuang Chen, Ying Wei, Jonathan Richard Schwarz

09:33 - 09:44

A Sample Offline Saves Time: Knowledge Distillation in the LLM Era
Anshumann, Mohd Abbas Zaidi, Akhil Kedia, Jinwoo Ahn, Taehwak Kwon, Kangwook Lee, Haejun Lee, Joohyung Lee
Anshumann, Mohd Abbas Zaidi, Akhil Kedia, Jinwoo Ahn, Taehwak Kwon, Kangwook Lee, Haejun Lee, Joohyung Lee

09:44 - 09:55

Cramming 1568 Tokens into a Single Vector and Back Again: Exploring the Limits of Embedding Space Capacity
Yuri Kuratov, Mikhail Arkhipov, Aydar Bulatov, Mikhail Burtsev
Yuri Kuratov, Mikhail Arkhipov, Aydar Bulatov, Mikhail Burtsev

09:55 - 10:06

SAD-LM: A Large-Scale Generalist Diffusion Language Model
Jaesung Tae, Hamish Ivison, Sachin Kumar, Arman Cohan
Jaesung Tae, Hamish Ivison, Sachin Kumar, Arman Cohan

10:06 - 10:17

Keys to Robust Edits: From Theoretical Insights to Practical Advances
Jianhao Yan, Futing Wang, Yun Luo, Yafu Li, Yue Zhang
Jianhao Yan, Futing Wang, Yun Luo, Yafu Li, Yue Zhang

10:17 - 10:28

Theoretical Analysis of Hierarchical Language Recognition and Generation by Transformers without Positional Encoding
Daichi Hayakawa, Issei Sato
Daichi Hayakawa, Issei Sato

Session 11: Multilinguality and Language Diversity (Wednesday 30 July, 09:00-10:30, Hall B)

09:22 - 09:33

MPO: Multilingual Safety Alignment via Reward Gap Optimization

Weixiang Zhao, Yulin Hu, Yang Deng, Tongtong Wu, Wenxuan Zhang, Jiahe Guo, An Zhang, Yanyan Zhao, Bing Qin, Tat-Seng Chua, Ting Liu

Weixiang Zhao, Yulin Hu, Yang Deng, Tongtong Wu, Wenxuan Zhang, Jiahe Guo, An Zhang, Yanyan Zhao, Bing Qin, Tat-Seng Chua, Ting Liu

Session 11: Multilinguality and Language Diversity (Wednesday 30 July, 09:00-10:30, Hall N.2)

09:00 - 09:11

Beyond N-Grams: Rethinking Evaluation Metrics and Strategies for Multilingual Abstractive Summarization

Itai Mondshine, Tzuf Paz-Argaman, Reut Tsarfaty

Itai Mondshine, Tzuf Paz-Argaman, Reut Tsarfaty

09:11 - 09:22

Bridging the Language Gaps in Large Language Models with Inference-Time Cross-Lingual Intervention

Weixuan Wang, Minghao Wu, Barry Haddow, Alexandra Birch

Weixuan Wang, Minghao Wu, Barry Haddow, Alexandra Birch

09:22 - 09:33

LangSAMP: Language-Script Aware Multilingual Pretraining

Yihong Liu, Haotian Ye, Chunlan Ma, Mingyang Wang, Hinrich Schuetze

Yihong Liu, Haotian Ye, Chunlan Ma, Mingyang Wang, Hinrich Schuetze

09:33 - 09:44

Second Language (Arabic) Acquisition of LLMs via Progressive Vocabulary Expansion

Jianqing Zhu, Huang Huang, Zhihang Lin, Juhao Liang, Zhengyang Tang, Khalid Almubarak, Mosen Alharthi, Bang An, Juncai He, Xiangbo Wu, Fei Yu, Junying Chen, MA Zhuoheng, Yuhao Du, He Zhang, Saied Alshahrani, Emad A. Alghamdi, Lian Zhang, Ruoyu Sun, Haizhou Li, Benyou Wang, Jinchao Xu

Jianqing Zhu, Huang Huang, Zhihang Lin, Juhao Liang, Zhengyang Tang, Khalid Almubarak, Mosen Alharthi, Bang An, Juncai He, Xiangbo Wu, Fei Yu, Junying Chen, MA Zhuoheng, Yuhao Du, He Zhang, Saied Alshahrani, Emad A. Alghamdi, Lian Zhang, Ruoyu Sun, Haizhou Li, Benyou Wang, Jinchao Xu

09:44 - 09:55

Building Better: Avoiding Pitfalls in Developing Language Resources when Data is Scarce

Nedjma Ousidhoum, Meriem Beloucif, Saif M. Mohammad

Nedjma Ousidhoum, Meriem Beloucif, Saif M. Mohammad

Session 11: Multimodality and Language Grounding to Vision, Robotics and Beyond (Wednesday 30 July, 09:00-10:30, Hall M.2)

09:00 - 09:11

Modeling Uncertainty in Composed Image Retrieval via Probabilistic Embedding

Haomiao Tang, Jinpeng Wang, Yuang Peng, GuangHao Meng, Ruisheng Luo, Bin Chen, Long Chen, Yaowei Wang, Shu-Tao Xia

Haomiao Tang, Jinpeng Wang, Yuang Peng, GuangHao Meng, Ruisheng Luo, Bin Chen, Long Chen, Yaowei Wang, Shu-Tao Xia

09:11 - 09:22

Centurio: On Drivers of Multilingual Ability of Large Vision-Language Model

Gregor Geigle, Florian Schneider, Carolin Holtermann, Chris Biemann, Radu Timofte, Anne Lauscher, Goran Glavaš

Gregor Geigle, Florian Schneider, Carolin Holtermann, Chris Biemann, Radu Timofte, Anne Lauscher, Goran Glavaš

09:22 - 09:33

FloorPlan-LLaMa: Aligning Architects' Feedback and Domain Knowledge in Architectural Floor Plan Generation

Jun Yin, Pengyu Zeng, Haoyuan Sun, Yuqin Dai, Han Zheng, Miao Zhang, Yachao Zhang, Shuai Lu

Jun Yin, Pengyu Zeng, Haoyuan Sun, Yuqin Dai, Han Zheng, Miao Zhang, Yachao Zhang, Shuai Lu

09:33 - 09:44

TheoremExplainAgent: Towards Video-based Multimodal Explanations for LLM Theorem Understanding

Max Ku, Thomas Chong, Jonathan Leung, Krish Shah, Alvin Yu, Wenhu Chen

Max Ku, Thomas Chong, Jonathan Leung, Krish Shah, Alvin Yu, Wenhu Chen

09:44 - 09:55

Enhancing Spoken Discourse Modeling in Language Models Using Gestural Cues

Varsha Suresh, M. Hamza Mughal, Christian Theobalt, Vera Demberg

Varsha Suresh, M. Hamza Mughal, Christian Theobalt, Vera Demberg

09:55 - 10:06

Retrospective Learning from Interactions

Zizhao Chen, Mustafa Omer Gul, Yiwei Chen, Gloria Geng, Anne Wu, Yoav Artzi

Zizhao Chen, Mustafa Omer Gul, Yiwei Chen, Gloria Geng, Anne Wu, Yoav Artzi

10:06 - 10:17

SHuBERT: Self-Supervised Sign Language Representation Learning via Multi-Stream Cluster Prediction

Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, Karen Livescu, Alexander H. Liu

Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, Karen Livescu, Alexander H. Liu

10:17 - 10:28

Diffusion Models Through a Global Lens: Are They Culturally Inclusive?

Zahra Bayramli, Ayhan Suleymanzade, Na Min An, Huzama Ahmad, Eunsu Kim, Junyeong Park, James Thorne, Alice Oh

Zahra Bayramli, Ayhan Suleymanzade, Na Min An, Huzama Ahmad, Eunsu Kim, Junyeong Park, James Thorne, Alice Oh

Session 11: NLP Applications (Wednesday 30 July, 09:00-10:30, Hall M.1)

09:00 - 09:11

Aristotle: Mastering Logical Reasoning with A Logic-Complete Decompose-Search-Resolve Framework

Jundong Xu, Hao Fei, Meng Luo, Qian Liu, Liangming Pan, William Yang Wang, Preslav Nakov, Mong-Li Lee, Wynne Hsu

Jundong Xu, Hao Fei, Meng Luo, Qian Liu, Liangming Pan, William Yang Wang, Preslav Nakov, Mong-Li Lee, Wynne Hsu

09:11 - 09:22

Toward Automatic Discovery of a Canine Phonetic Alphabet

Theron S. Wang, Xingyuan Li, Hridayesh Lekhak, Tuan Minh Dang, Mengyue Wu, Kenny Q. Zhu

Theron S. Wang, Xingyuan Li, Hridayesh Lekhak, Tuan Minh Dang, Mengyue Wu, Kenny Q. Zhu

09:22 - 09:33

ReflecTool: Towards Reflection-Aware Tool-Augmented Clinical Agents

Yusheng Liao, Shuyang Jiang, Yanfeng Wang, Yu Wang

Yusheng Liao, Shuyang Jiang, Yanfeng Wang, Yu Wang

09:33 - 09:44

A Troublemaker with Contagious Jailbreak Makes Chaos in Honest Towns

Tianyi Men, Pengfei Cao, Zhuoran Jin, Yubo Chen, Kang Liu, Jun Zhao

Tianyi Men, Pengfei Cao, Zhuoran Jin, Yubo Chen, Kang Liu, Jun Zhao

09:44 - 09:55

RoToR: Towards More Reliable Responses for Order-Invariant Inputs

Soyoung Yoon, Dongha Ahn, Youngwon Lee, Minkyu Jung, HyungJoo Jang, seung-won hwang

Soyoung Yoon, Dongha Ahn, Youngwon Lee, Minkyu Jung, HyungJoo Jang, seung-won hwang

09:55 - 10:06

Enhancing Mathematical Reasoning in LLMs by Stepwise Correction

Zhenyu Wu, Qingkai Zeng, Zhihan Zhang, Zhaoxuan Tan, Chao Shen, Meng Jiang

Zhenyu Wu, Qingkai Zeng, Zhihan Zhang, Zhaoxuan Tan, Chao Shen, Meng Jiang

10:06 - 10:17

XDAC: XAI-Driven Detection and Attribution of LLM-Generated News Comments in Korean

Wooyoung Go, Hyoungshick Kim, Alice Oh, Yongdae Kim

Wooyoung Go, Hyoungshick Kim, Alice Oh, Yongdae Kim

10:17 - 10:28

Knowledge Tracing in Programming Education Integrating Students' Questions

Doyoun Kim, Suin Kim, Yohan Jo

Doyoun Kim, Suin Kim, Yohan Jo

Session 11: Question Answering (Wednesday 30 July, 09:00-10:30, Room 1.86)

09:00 - 09:11

Which of These Best Describes Multiple Choice Evaluation with LLMs? A) Forced B) Flawed C) Fixable D) All of the Above

Nishant Balepur, Rachel Rudinger, Jordan Lee Boyd-Graber

Nishant Balepur, Rachel Rudinger, Jordan Lee Boyd-Graber

09:11 - 09:22

Knowledge Boundary of Large Language Models: A Survey

Moxin Li, Yong Zhao, Wenxuan Zhang, Shuaiyi Li, Wenya Xie, See-Kiong Ng, Tat-Seng Chua, Yang Deng

Moxin Li, Yong Zhao, Wenxuan Zhang, Shuaiyi Li, Wenya Xie, See-Kiong Ng, Tat-Seng Chua, Yang Deng

09:22 - 09:33

Shifting from Ranking to Set Selection for Retrieval Augmented Generation

Dahyun Lee, Yongrae Jo, Haeju Park, Moontae Lee

Dahyun Lee, Yongrae Jo, Haeju Park, Moontae Lee

09:33 - 09:44

Masking in Multi-hop QA: An Analysis of How Language Models Perform with Context Permutation

Wenyu Huang, Pavlos Vougiouklis, Mirella Lapata, Jeff Z. Pan

Wenyu Huang, Pavlos Vougiouklis, Mirella Lapata, Jeff Z. Pan

09:44 - 09:55

The Distracting Effect: Understanding Irrelevant Passages in RAG

Chen Amiraz, Florin Cuconasu, Simone Filice, Zohar Karnin

Chen Amiraz, Florin Cuconasu, Simone Filice, Zohar Karnin

09:55 - 10:06

Alleviating Hallucinations from Knowledge Misalignment in Large Language Models via Selective Abstention Learning

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuchun Fan, Xiachong Feng, Yuxuan Gu, Yangfan Ye, Liang Zhao, Weihong Zhong, Baoxin Wang, Dayong Wu, Guoping Hu, Lingpeng Kong, Tong Xiao, Ting Liu, Bing Qin

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuchun Fan, Xiachong Feng, Yuxuan Gu, Yangfan Ye, Liang Zhao, Weihong Zhong, Baoxin Wang, Dayong Wu, Guoping Hu, Lingpeng Kong, Tong Xiao, Ting Liu, Bing Qin

10:06 - 10:17

SeaKR: Self-aware Knowledge Retrieval for Adaptive Retrieval Augmented Generation

Zijun Yao, Weijian Qi, Liangming Pan, Shulin Cao, Linmei Hu, Liu Weichuan, Lei Hou, Juanzi Li

Zijun Yao, Weijian Qi, Liangming Pan, Shulin Cao, Linmei Hu, Liu Weichuan, Lei Hou, Juanzi Li

10:17 - 10:28

Benchmarking Video-Language Models for Embodied Motion Cognition in Urban Open-Ended Spaces

Baining Zhao, Jianjie Fang, Zichao Dai, Ziyou Wang, Jirong Zha, Weichen Zhang, Chen Gao, Yue Wang, Jinqiang Cui, Xinlei Chen, Yong Li

Baining Zhao, Jianjie Fang, Zichao Dai, Ziyou Wang, Jirong Zha, Weichen Zhang, Chen Gao, Yue Wang, Jinqiang Cui, Xinlei Chen, Yong Li

Session 11: Resources and Evaluation (Wednesday 30 July, 09:00-10:30, Hall C)

09:22 - 09:33

Palm: A Culturally Inclusive and Linguistically Diverse Dataset for Arabic LLMs

Fakhraddin Alwajih, Abdellah EL MEKKI, Samar Mohamed Magdy, AbdelRahim A. Elmadany, OMER NACAR, El Moatez Billah Nagoudi, Reem Abdel-Salam, Hanin atwany, Youssef Nafea, Abdulfattah Mohammed Yahya, Rahaf Alhamouri, Hamzah A. Alsayadi, Hiba Zayed, Sara Shatnawi, Serry Sibaee, Yasir ECH-CHAMMAKHY, Walid Al-Dhabyani, Marwa Mohamed Ali, Imen JARRAYA, Ahmed Oumar El-Shangiti, Aisha Alraeesi, Mohammed Anwar AL-Ghrawi, Abdulrahman S. Al-Batati, Elgizouli Mohamed, Noha Taha Elgindi, Muhammed Saeed, Houdaifa Atou, Issam AIT YAHIA, Abdelhak Bouayad, Mohammed Machrouh, AMAL MAKOUAR, Dania Alkawi, Mukhtar Mohamed, Safaa Taher Abdelfadil, Amine Ziad Ounnoughene, Anfel ROUABHIA, Rwaa Assi, Ahmed Sorkatti, Mohamedou cheikh tourad, Anis Koubaa, Ismail Berrada, Mustafa Jarrar, Shady Shehata, Muhammad Abdul-Mageed

Fakhraddin Alwajih, Abdellah EL MEKKI, Samar Mohamed Magdy, AbdelRahim A. Elmadany, OMER NACAR, El Moatez Billah Nagoudi, Reem Abdel-Salam, Hanin atwany, Youssef Nafea, Abdulfattah Mohammed Yahya, Rahaf Alhamouri, Hamzah A. Alsayadi, Hiba Zayed, Sara Shatnawi, Serry Sibaee, Yasir ECH-CHAMMAKHY, Walid Al-Dhabyani, Marwa Mohamed Ali, Imen JARRAYA, Ahmed Oumar El-Shangiti, Aisha Alraeesi, Mohammed Anwar AL-Ghrawi, Abdulrahman S. Al-Batati, Elgizouli Mohamed, Noha Taha Elgindi, Muhammed Saeed, Houdaifa Atou, Issam AIT YAHIA, Abdelhak Bouayad, Mohammed Machrouh, AMAL MAKOUAR, Dania Alkawi, Mukhtar Mohamed, Safaa Taher Abdelfadil, Amine Ziad Ounnoughene, Anfel ROUABHIA, Rwaa Assi, Ahmed Sorkatti, Mohamedou cheikh tourad, Anis Koubaa, Ismail Berrada, Mustafa Jarrar, Shady Shehata, Muhammad Abdul-Mageed

Session 11: Special Theme: Generalization of NLP Models (Wednesday 30 July, 09:00-10:30, Hall A)

09:00 - 09:11

Between Circuits and Chomsky: Pre-pretraining on Formal Languages Imparts Linguistic Biases

Michael Y. Hu, Jackson Petty, Chuan Shi, William Merrill, Tal Linzen

Michael Y. Hu, Jackson Petty, Chuan Shi, William Merrill, Tal Linzen

09:11 - 09:22

Revisiting Compositional Generalization Capability of Large Language Models Considering Instruction Following Ability

Yusuke Sakai, Hidetaka Kamigaito, Taro Watanabe

Yusuke Sakai, Hidetaka Kamigaito, Taro Watanabe

09:22 - 09:33

Behavioural vs. Representational Systematicity in End-to-End Models: An Opinionated Survey

Ivan Vegner, Sydelle de Souza, Valentin Forch, Martha Lewis, Leonidas A. A. Doumas

Ivan Vegner, Sydelle de Souza, Valentin Forch, Martha Lewis, Leonidas A. A. Doumas

09:33 - 09:44

Steering off Course: Reliability Challenges in Steering Language Models

Patrick Queiroz Da Silva, Hari Sethuraman, Dheeraj Rajagopal, Hannaneh Hajishirzi, Sachin Kumar

Patrick Queiroz Da Silva, Hari Sethuraman, Dheeraj Rajagopal, Hannaneh Hajishirzi, Sachin Kumar

Session 11: TACL Journal (Wednesday 30 July, 09:00-10:30, Room 1.15-16)

10:17 - 10:28

Know Your Limits: A Survey of Abstention in Large Language Models

Bingbing Wen Jihan Yao Shangbin Feng Chenjun Xu Yulia Tsvetkov Bill Howe Lucy Lu Wang

Abstention, the refusal of large language models (LLMs) to provide an answer, is increasingly recognized for its potential to mitigate hallucinations and enhance safety in LLM systems. In this survey, we introduce a framework to examine abstention from three perspectives: the query, the model, and human values. We organize the literature on abstention methods, benchmarks, and evaluation metrics using this framework, and discuss merits and limitations of prior work. We further identify and motivate areas for future work, centered around whether abstention can be achieved as a meta-capability that transcends specific tasks or domains, while still providing opportunities to optimize abstention abilities based on context.

Poster Presentation

Session 12: Computational Social Science and Cultural Analytics (Wednesday 30 July, 11:00-12:30, Hall 4/5)

Do Large Language Models Know Folktales? A Case Study of Yokai in Japanese Folktales

Ayuto Tsutsumi, Yuu Jinnai

Although Large Language Models (LLMs) have demonstrated strong language understanding and generation abilities across various languages, their cultural knowledge is often limited to English-speaking communities, which can marginalize the cultures of non-English communities. To address the problem, the evaluation of the cultural awareness of the LLMs and the methods to develop culturally aware LLMs have been investigated. In this study, we focus on evaluating knowledge of folktales, a key medium for conveying and circulating culture. In particular, we focus on Japanese folktales, specifically on knowledge of Yokai. Yokai are supernatural creatures originating from Japanese folktales that continue to be popular motifs in art and entertainment today. Yokai have long served as a medium for cultural expression, making them an ideal subject for assessing the cultural awareness of LLMs. We introduce YokaiEval, a benchmark dataset consisting of 809 multiple-choice questions (each with four options) designed to probe knowledge about yokai. We evaluate the performance of 31 Japanese and multilingual LLMs on this dataset. The results show that models trained with Japanese language resources achieve higher accuracy than English-centric models, with those that underwent continued pretraining in Japanese, particularly those based on Llama-3, performing especially well.

BEDAA: Bayesian Enhanced DeBERTa for Uncertainty-Aware Authorship Attribution

Iqra Zahid, Youcheng Sun, Riza Batista-Navarro

Authorship Attribution (AA) seeks to identify the author of a given text, yet existing methods often struggle with trustworthiness and interpretability, particularly across different domains, languages, and stylistic variations. These challenges arise from the absence of uncertainty quantification and the inability of current models to adapt to diverse authorship tasks. To address these limitations, we introduce BEDAA, a Bayesian-Enhanced DeBERTa framework that integrates Bayesian reasoning with transformer-based language models to enable uncertainty-aware and interpretable authorship attribution. BEDAA achieves up to 19.69% improvement in F1-score across multiple authorship attribution tasks, including binary, multiclass, and dynamic authorship detection. By incorporating confidence ranking, uncertainty decomposition, and probabilistic reasoning, BEDAA improves robustness while offering transparent decision-making processes. Furthermore, BEDAA extends beyond traditional AA by demonstrating its effectiveness in human vs. machine-generated text classification, code authorship detection, and cross-lingual attribution. These advances establish BEDAA as a generalised, interpretable, and adaptable framework for modern authorship attribution challenges.

Mind the (Belief) Gap: Group Identity in the World of LLMs

Angana Borah, Marwa Houalla, Rada Mihalcea

Social biases and belief-driven behaviors can significantly impact Large Language Models' (LLMs') decisions on several tasks. As LLMs are increasingly used in multi-agent systems for societal simulations, their ability to model fundamental group psychological characteristics remains critical yet under-explored. In this study, we present a multi-agent framework that simulates belief congruence, a classical group psychology theory that plays a crucial role in shaping societal interactions and preferences. Our findings reveal that LLMs exhibit amplified belief congruence compared to humans, across diverse contexts. We further investigate the implications of this behavior on two downstream tasks: (1) misinformation dissemination and (2) LLM learning, finding that belief congruence in LLMs increases misinformation dissemination and impedes learning. To mitigate these negative impacts, we propose strategies inspired by: (1) contact hypothesis, (2) accuracy nudges, and (3) global citizenship framework. Our results show that the best strategies reduce misinformation dissemination by up to (37%) and enhance learning by (11%). Bridging social psychology and AI, our work provides insights to navigate real-world interactions using LLMs while addressing belief-driven biases.

The Impact of Large Language Models in Academia: from Writing to Speaking

Mingmeng Geng, Caixi Chen, Yanru Wu, Yao Wan, Pan Zhou, Dongping Chen

Large language models (LLMs) are increasingly impacting human society, particularly in textual information. Based on more than 30,000 papers and 1,000 presentations from machine learning conferences, we examined and compared the words used in writing and speaking, representing the first large-scale study of how LLMs influence the two main modes of verbal communication and expression within the same group of people. Our empirical results show that LLM-style words such as significant have been used more frequently in abstracts and oral presentations. The implicit impact on human expression like writing and speaking is beginning to emerge and is likely to grow in the future. We take the first step in building an automated monitoring platform to record its longitudinal changes to call attention to the implicit influence and ripple effect of LLMs on human society.

LLMTaxo: Leveraging Large Language Models for Constructing Taxonomy of Factual Claims from Social Media

Haiqi Zhang, Zhengyuan Zhu, Zeyu Zhang, Chengkai Li

With the vast expansion of content on social media platforms, analyzing and comprehending online discourse has become increasingly complex. This paper introduces LLMTaxo, a novel framework leveraging large language models for the automated construction of taxonomy of factual claims from social media by generating topics from multi-level granularities. This approach aids stakeholders in more effectively navigating the social media landscapes. We implement this framework with different models across three distinct datasets and introduce specially designed taxonomy evaluation metrics for a comprehensive assessment. With the evaluations from both human evaluators and GPT-4, the results indicate that LLMTaxo effectively categorizes factual claims from social media, and reveals that certain models perform better on specific datasets.

How Personality Traits Shape LLM Risk-Taking Behaviour

John Hartley, Conor Brian Hamill, Dale Seddon, Devesh Batra, Ramin Okhrati, Raad Khraishi

Large Language Models (LLMs) are increasingly deployed as autonomous agents for simulation and decision-making, necessitating a deeper understanding of their decision-making behaviour under risk. We investigate the relationship between LLMs' personality traits and risk-propensity, applying Cumulative Prospect Theory (CPT) and the Big Five personality framework. We compare the behaviour of several LLMs to human baselines. Our findings show that the majority of the models investigated are risk-neutral rational agents, whilst displaying higher Conscientiousness and Agreeableness traits, coupled with lower Neuroticism. Interventions on Big Five traits, particularly Openness, influence the risk-propensity of several LLMs. Advanced models mirror human personality-risk patterns, suggesting that cognitive biases can be surfaced by optimal prompting. However, their distilled variants show no cognitive bias, suggesting limitations to knowledge transfer processes. Notably, Openness emerges as the most influential factor to risk-propensity, aligning with human baselines. In contrast, less advanced models demonstrate inconsistent generalization of the personality-risk relationship. This research advances our understanding of LLM behaviour under risk and highlights the potential and limitations of personality-based interventions in shaping LLM decision-making.

Does Rationale Quality Matter? Enhancing Mental Disorder Detection via Selective Reasoning Distillation

Hoyun Song, Huije Lee, Jisu Shin, Sukmin Cho, Changgeon Ko, Jong C. Park

The detection of mental health problems from social media and the interpretation of these results have been extensively explored. Research has shown that incorporating clinical symptom information into a model enhances domain expertise, improving its detection and interpretation performance. While large language models (LLMs) are shown to be effective for generating explanatory rationales in mental health detection, their substantially big parameter size and high computational cost limit their practicality. Reasoning distillation transfers this ability to smaller language models (SLMs), but inconsistencies in the relevance and domain alignment of LLM-generated rationales pose a challenge. This paper investigates how rationale quality impacts SLM performance in mental health detection and explanation generation. We hypothesize that ensuring high-quality and domain-relevant rationales enhances the distillation. To this end, we propose a framework that selects rationales based on their alignment with expert clinical reasoning. Experiments show that our quality-focused approach significantly enhances SLM performance in both mental disorder detection and rationale generation. This work highlights the importance of rationale quality and offers an insightful framework for knowledge transfer in mental health applications.

Robustness and Confounders in the Demographic Alignment of LLMs with Human Perceptions of Offensiveness

Shayan Alipour, Indira Sen, Mattia Samory, Tanu Mitra

Despite a growing literature finding that large language models (LLMs) exhibit demographic biases, reports with whom they align best are hard to generalize or even contradictory. In this work, we examine the alignment of LLMs with human annotations in five offensive language datasets, comprising approximately 220K annotations. While demographic traits, particularly race, influence alignment, these effects vary across datasets and are often entangled with other factors. Confounders introduced in the annotation process\—such as document difficulty, annotator sensitivity, and within-group agreement\—account for more variation in alignment patterns than demographic traits. Alignment increases with annotator sensitivity and group agreement, and decreases with document difficulty. Our results underscore the importance of multi-dataset analyses and confounder-aware methodologies in developing robust measures of demographic bias.

Revealing Hidden Mechanisms of Cross-Country Content Moderation with Natural Language Processing

Neemesh Yadav, Jiarui Liu, Francesco Ortu, Roya Ensafi, Zhijing Jin, Rada Mihalcea

The ability of Natural Language Processing (NLP) methods to categorize text into multiple classes has motivated their use in online content moderation tasks, such as hate speech and fake news detection. However, there is limited understanding of how or why these methods make such decisions, or why certain content is moderated in the first place. To investigate the hidden mechanisms behind content moderation, we explore multiple directions: 1) training classifiers to reverse-engineer content moderation decisions across countries; 2) explaining content moderation decisions by analyzing Shapley values and LLM-guided explanations. Our primary focus is on content moderation decisions made across countries, using pre-existing corpora sampled from the Twitter Stream Grab. Our experiments reveal interesting patterns in censored posts, both across countries and over time. Through human evaluations of LLM-generated explanations across three LLMs, we assess the effectiveness of using LLMs in content moderation. Finally, we discuss potential future directions, as well as the limitations and ethical considerations of this work.

Tell, Don't Show: Leveraging Language Models' Abstractive Retellings to Model Literary Themes

Li Lucy, Camilla Griffiths, Sarah Levine, Jennifer L Eberhardt, Dorottya Demszky, David Bamman

Conventional bag-of-words approaches for topic modeling, like latent Dirichlet allocation (LDA), struggle with literary text. Literature challenges lexical methods because narrative language focuses on immersive sensory details instead of abstractive description or exposition: writers are advised to *show, don't tell*. We propose Retell, a simple, accessible topic modeling approach for literature. Here, we prompt resource-efficient, generative language models (LMs) to *tell* what passages *show*, thereby translating narratives' surface forms into higher-level concepts and themes. By running LDA on LMs' retellings of passages, we can obtain more precise and informative topics than by running LDA alone or by directly asking LMs to list topics. To investigate the potential of our method for cultural analytics, we compare our method's outputs to expert-guided annotations in a case study on racial/cultural identity in high school English language arts books.

Just KIDDIN' : Knowledge Infusion and Distillation for Detection of INdecent Memes

Rahul Garg, Trilok Padhi, Hemang Jain, Ugur Kursuncu, Ponnurangam Kumaraguru

Detecting toxicity in online multimodal environments such as memes remains a challenging task due to the complex contextual connections across modalities (e.g., text and visual), demanding both common sense reasoning and contextual awareness. To bridge this gap, we propose a hybrid neurosymbolic framework that unifies (1) distillation of implicit contextual knowledge (e.g., sarcasm, cultural references) from Large Vision-Language Models (LVLMs) and (2) infusion of explicit relational semantics through sub-graphs from Knowledge Graphs (KGs). Experimental results on two benchmark datasets show the superior performance of our approach, Knowledge-Infused Distilled Vision-Language Model (KID-VLM), over the state-of-the-art baselines across AUC, and F1, with improvements of 0.5%, and 10.6%, respectively in HatefulMemes Benchmark across variants. Further, KID-VLM demonstrates better generalizability and achieves best performance across all baselines in HarMeme Dataset with an 6.3% and 3.2% in F1 and AUC. Given the contextual complexity of the toxicity detection task, KID-VLM showcases the significance of learning compact models (500M parameters) from both explicit (i.e., KG) and implicit (i.e., LVLMs) contextual cues incorporated through a hybrid neurosymbolic approach. KID-VLM provides an accurate and scalable approach to recognize toxic content for safer online environments in real-world applications. Our code is available at Anonymous Github at: https://anonymous.4open.science/r/MultimodalToxicity-124F/README.md.

When AI Writes Like Humans: Capturing the Emergent Patterns of Literary Judgment via Intrinsic Textual Metrics

Guillermo Marco, Julio Gonzalo, Víctor Fresno, Juan-Luis Suárez

As Large Language Models (LLMs) increasingly produce human-quality text, understanding how readers perceive and evaluate AI-generated literature becomes critical. This paper investigates the interplay between measurable textual features and reader preferences in the context of short stories. We analyze multiple datasets where readers (ranging from literary experts to lay audiences) evaluated both human-authored and LLM-generated texts. We propose to model reader's preferences as weighted combinations of intrinsic textual features such as coherence, sentiment volatility, thematic entropy, etc. Results show that (1) when LLM-generated text closely approximates the feature distributions of human text, readers struggle to distinguish them; (2) expert and non-expert readers prioritize distinctly different textual features, revealing systematic variations in literary judgment; and (3) these preference patterns are remarkably stable across diverse datasets. Our framework provides a step towards understanding and potentially tailoring AI-generated literature to meet the expectations of specific audiences.

Just Put a Human in the Loop? Investigating LLM-Assisted Annotation for Subjective Tasks

Hope Schroeder, Deb Roy, Jad Kabbara

LLM use in annotation is becoming widespread, and given LLMs' overall promising performance and speed, putting humans in the loop to simply "review" LLM annotations can be tempting. In subjective tasks with multiple plausible answers, this can impact both evaluation of LLM performance, and analysis using these labels in a social science task downstream. In a pre-registered experiment with 350 unique annotators and 7,000 annotations across 4 conditions, 2 models, and 2 datasets, we find that presenting crowdworkers with LLM-generated annotation suggestions did not make them faster annotators, but did improve their self-reported confidence in the task. More importantly, annotators strongly took the LLM suggestions, significantly changing the label distribution compared to the baseline. We show that when these labels created with LLM assistance are used to evaluate LLM performance, reported model performance significantly increases. We show how changes in label distributions as a result of LLM assistance can affect conclusions drawn by analyzing even "human-approved" LLM-annotated datasets. We believe our work underlines the importance of understanding the impact of LLM-assisted annotation on subjective, qualitative tasks, on the creation of gold data for training and testing, and on the evaluation of NLP systems on subjective tasks.

MIRAGE: Exploring How Large Language Models Perform in Complex Social Interactive Environments

Yin Cai, Zhouhong Gu, Zhaohan Du, Zheyu Ye, Shaosheng Cao, Yiqian xu, Hongwei Feng, Ping Chen

Large Language Models (LLMs) have shown remarkable capabilities in environmental perception, reasoning-based decision-making, and simulating complex human behaviors, particularly in interactive role-playing contexts. This paper introduces the Multiverse Interactive Role-play Ability General Evaluation (MIRAGE), a comprehensive framework designed to assess LLMs' proficiency in portraying advanced human behaviors through murder mystery games. MIRAGE features eight intricately crafted scripts encompassing diverse themes and styles, providing a rich simulation. To evaluate LLMs' performance, MIRAGE employs four distinct methods: the Trust Inclination Index (TII) to measure dynamics of trust and suspicion, the Clue Investigation Capability (CIC) to measure LLMs' capability of conducting information, the Interactivity Capability Index (ICI) to assess role-playing capabilities and the Script Compliance Index (SCI) to assess LLMs' capability of understanding and following instructions. Our experiments indicate that even popular models like GPT-4 face significant challenges in navigating the complexities presented by the MIRAGE. The datasets and simulation codes are available in https://github.com/lime728/MIRAGE.

Capturing Author Self Beliefs in Social Media Language

Siddharth Mangalik, Adithya V Ganesan, Abigail B. Wheeler, Nicholas Kerry, Jeremy D. W. Clifton, H. Schwartz, Ryan L. Boyd

Measuring the prevalence and dimensions of self beliefs is essential for understanding human self-perception and various psychological outcomes. In this paper, we develop a novel task for classifying language that contains explicit or implicit mentions of the author's self beliefs. We contribute a set of 2,000 human-annotated self beliefs, 100,000 LLM-labeled examples, and 10,000 surveyed self belief paragraphs. We then evaluate several encoder-based classifiers and training routines for this task. Our trained model, SelfAwareNet, achieved an AUC of 0.944, outperforming 0.839 from OpenAI's state-of-the-art GPT-4o model. Using this model we derive data-driven categories of self beliefs and demonstrate their ability to predict valence, depression, anxiety, and stress. We release the resulting self belief classification model and annotated datasets for use in future research.

ECERC: Evidence-Cause Attention Network for Multi-Modal Emotion Recognition in Conversation

Tao Zhang, Zhenhua Tan

Multi-modal Emotion Recognition in Conversation (MMERC) aims to identify speakers' emotional states using multi-modal conversational data, significant for various domains. MMERC requires addressing emotional causes: contextual factors that influence emotions, alongside emotional evidence directly expressed in the target utterance. Existing methods primarily model general conversational dependencies, such as sequential utterance relationships or inter-speaker dynamics, but fall short in capturing diverse and detailed emotional causes, including emotional contagion, influences from others, and self-referenced or externally introduced events. To address these limitations, we propose the Evidence-Cause Attention Network for Multi-Modal Emotion Recognition in Conversation (ECERC). ECERC integrates emotional evidence with contextual causes through five stages: Evidence Gating extracts and refines emotional evidence across modalities; Cause Encoding captures causes from conversational context; Evidence-Cause Interaction uses attention to integrate evidence with diverse causes, generating rich candidate features for emotion inference; Feature Gating adaptively weights contributions of candidate features; and Emotion Classification classifies emotions. We evaluate ECERC on two widely used benchmark datasets, IEMOCAP and MELD. Experimental results show that ECERC achieves competitive performance in weighted F1-score and accuracy, demonstrating its effectiveness in MMERC

Hierarchical Level-Wise News Article Clustering via Multilingual Matryoshka Embeddings

Hans William Alexander Hanley, Zakir Durumeric

Contextual large language model embeddings are increasingly utilized for topic modeling and clustering. However, current methods often scale poorly, rely on opaque similarity metrics, and struggle in multilingual settings. In this work, we present a novel, scalable, interpretable, hierarchical, and multilingual approach to clustering news articles and social media data. To do this, we first train multilingual Matryoshka embeddings that can determine story similarity at varying levels of granularity based on which subset of the dimensions of the embeddings is examined. This embedding model achieves state-of-the-art performance on the SemEval 2022 Task 8 test dataset (Pearson \rho = 0.816). Once trained, we develop an efficient hierarchical clustering algorithm that leverages the hierarchical nature of Matryoshka embeddings to identify unique news stories, narratives, and themes. We conclude by illustrating how our approach can identify and cluster stories, narratives, and overarching themes within real-world news datasets.

Making FETCH! Happen: Finding Emergent Dog Whistles Through Common Habitats

Kuleen Sasse, Carlos Alejandro Aguirre, Isabel Cachola, Sharon Levy, Mark Dredze

Dog whistles are coded expressions with dual meanings: one intended for the general public (outgroup) and another that conveys a specific message to an intended audience (ingroup). Often, these expressions are used to convey controversial political opinions while maintaining plausible deniability and slip by content moderation filters. Identification of dog whistles relies on curated lexicons, which have trouble keeping up to date. We introduce FETCH!, a task for finding novel dog whistles in massive social media corpora. We find that state-of-the-art systems fail to achieve meaningful results across three distinct social media case studies. We present EarShot, a strong baseline system that combines the strengths of vector databases and Large Language Models (LLMs) to efficiently and effectively identify new dog whistles.

Tracking Life's Ups and Downs: Mining Life Events from Social Media Posts for Mental Health Analysis

Minghao Lv, Siyuan Chen, Haoan Jin, Minghao Yuan, Qianqian Ju, Yujia Peng, Kenny Q. Zhu, Mengyue Wu

Social media platforms possess considerable potential in the realm of exploring mental health. Previous research has indicated that major life events can greatly impact individuals' mental health. However, due to the complexity and ambiguity nature of life events, shedding its light on social media data is quite challenging. In this paper, we are dedicated to uncovering life events mentioned in posts on social media. We hereby provide a carefully-annotated social media event dataset, PsyEvent, which encompasses 12 major life event categories that are likely to occur in everyday life. This dataset is human-annotated under iterative procedure and boasts a high level of quality. Furthermore, by applying the life events extracted from posts to downstream tasks such as early risk detection of depression and suicide risk prediction, we have observed a considerable improvement in performance. This suggests that extracting life events from social media can be beneficial for the analysis of individuals' mental health.

When People are Floods: Analyzing Dehumanizing Metaphors in Immigration Discourse with Large Language Models

Julia Mendelsohn, Ceren Budak

Metaphor, discussing one concept in terms of another, is abundant in politics and can shape how people understand important issues. We develop a computational approach to measure metaphorical language, focusing on immigration discourse on social media. Grounded in qualitative social science research, we identify seven concepts evoked in immigration discourse (e.g. water or vermin). We propose and evaluate a novel technique that leverages both word-level and document-level signals to measure metaphor with respect to these concepts. We then study the relationship between metaphor, political ideology, and user engagement in 400K US tweets about immigration. While conservatives tend to use dehumanizing metaphors more than liberals, this effect varies widely across concepts. Moreover, creature-related metaphor is associated with more retweets, especially for liberal authors. Our work highlights the potential for computational methods to complement qualitative approaches in understanding subtle and implicit language in political discourse.

Conspiracy Theories and Where to Find Them on TikTok

Francesco Corso, Francesco Pierri, Gianmarco De Francisci Morales

TikTok has skyrocketed in popularity over recent years, especially among younger audiences. However, there are public concerns about the potential of this platform to promote and amplify harmful content. This study presents the first systematic analysis of conspiracy theories on TikTok. By leveraging the official TikTok Research API we collect a longitudinal dataset of 1.5M videos shared in the U.S. over three years. We estimate a lower bound on the prevalence of conspiratorial videos (up to 1000 new videos per month) and evaluate the effects of TikTok's Creativity Program for monetization, observing an overall increase in video duration regardless of content. Lastly, we evaluate the capabilities of state-of-the-art open-weight Large Language Models to identify conspiracy theories from audio transcriptions of videos. While these models achieve high precision in detecting harmful content (up to 96%), their overall performance remains comparable to fine-tuned traditional models such as RoBERTa. Our findings suggest that Large Language Models can serve as an effective tool for supporting content moderation strategies aimed at reducing the spread of harmful content on TikTok.

Literary Evidence Retrieval via Long-Context Language Models

Katherine Thai, Mohit Iyyer

How well do modern long-context language models understand literary fiction? We explore this question via the task of literary evidence retrieval, repurposing the RELiC dataset of Thai et al. (2022) to construct a benchmark where the entire text of a primary source (e.g., The Great Gatsby) is provided to an LLM alongside literary criticism with a missing quotation from that work. This setting, in which the model must generate the missing quotation, mirrors the human process of literary analysis by requiring models to perform both global narrative reasoning and close textual examination. We curate a high-quality subset of 292 examples through extensive filtering and human verification. Our experiments show that recent reasoning models, such as Gemini 2.5 Pro can exceed human expert performance (62.5% vs. 50% accuracy). In contrast, the best open-weight model achieves only 29.1% accuracy, highlighting a wide gap in interpretive reasoning between open and closed-weight models. Despite their speed and apparent accuracy, even the strongest models struggle with nuanced literary signals and overgeneration, signaling open challenges for applying LLMs to literary analysis. We release our dataset and evaluation code to encourage future work in this direction.

Empathy Prediction from Diverse Perspectives

Francine Chen, Scott Carter, Tatiana Lau, Nayeli Suseth Bravo, Sumanta Bhattacharyya, Kate Sieck, Charlene C. Wu

A person's perspective on a topic can influence their empathy towards a story. To investigate the use of personal perspective in empathy prediction, we collected a dataset, EmpathyFromPerspectives, where a user rates their empathy towards a story by a person with a different perspective on a prompted topic. We observed in the dataset that user perspective can be important for empathy prediction and developed a model, PPEP, that uses a rater's perspective as context for predicting the rater's empathy towards a story. Experiments comparing PPEP with baseline models show that use of personal perspective significantly improves performance. A user study indicated that human empathy ratings of stories generally agreed with PPEP's relative empathy rankings.

Culture Matters in Toxic Language Detection in Persian

Zahra Bokaei, Walid Magdy, Bonnie Webber

Toxic language detection is crucial for creating safer online environments and limiting the spread of harmful content. While toxic language detection has been under-explored in Persian, the current work compares different methods for this task, including fine-tuning, data enrichment, zero-shot and few-shot learning, and cross-lingual transfer learning. What is especially compelling is the impact of cultural context on transfer learning for this task: We show that the language of a country with cultural similarities to Persian yields better results in transfer learning. Conversely, the improvement is lower when the language comes from a culturally distinct country.

Towards Enhanced Immersion and Agency for LLM-based Interactive Drama

Hongqiu Wu, Weiqi Wu, Tianyang Xu, Jiameng Zhang, hai zhao

LLM-based Interactive Drama is a novel AI-based dialogue scenario, where the user (i.e. the player) plays the role of a character in the story, has conversations with characters played by LLM agents, and experiences an unfolding story. This paper begins with understanding interactive drama from two aspects: Immersion—the player's feeling of being present in the story—and Agency—the player's ability to influence the story world. Both are crucial to creating an enjoyable interactive experience, while they have been underexplored in previous work. To enhance these two aspects, we first propose Playwriting-guided Generation, a novel method that helps LLMs craft dramatic stories with substantially improved structures and narrative quality. Additionally, we introduce Plot-based Reflection for LLM agents to refine their reactions to align with the player's intentions. Our evaluation relies on human judgment to assess the gains of our methods in terms of immersion and agency.

LISTN: Lexicon induction with socio-temporal nuance

Christine de Kock

In-group language is an important signifier of group dynamics. This paper proposes a novel method for inducing lexicons of in-group language, which incorporates its socio-temporal context. Existing methods for lexicon induction do not capture the evolving nature of in-group language, nor the social structure of the community. Using dynamic word and user embeddings trained on conversations from online anti-women communities, our approach outperforms prior methods for lexicon induction. We develop a test set for the task of lexicon induction and a new lexicon of manosphere language, validated by human experts, which quantifies the relevance of each term to a specific sub-community at a given point in time. Finally, we present novel insights on in-group language which illustrate the utility of this approach.

Incongruity-aware Tension Field Network for Multi-modal Sarcasm Detection

Jiecheng Zhang, C.L.Philip Chen, Shuzhen Li, Tong Zhang

Multi-modal sarcasm detection (MSD) identifies sarcasm and accurately understands users' real attitudes from text-image pairs. Most MSD researches explore the incongruity of text-image pairs as sarcasm information through consistency preference methods. However, these methods prioritize consistency over incongruity and blur incongruity information under their global feature aggregation mechanisms, leading to incongruity distortions and model misinterpretations. To address the above issues, this paper proposes a pioneering inconsistency preference method called incongruity-aware tension field network (ITFNet) for multi-modal sarcasm detection tasks. Specifically, ITFNet extracts effective text-image feature pairs in fact and sentiment perspectives. It then constructs a fact/sentiment tension field with discrepancy metrics to capture the contextual tone and polarized incongruity after the iterative learning of tension intensity, effectively highlighting incongruity information during such inconsistency preference learning. It further standardizes the polarized incongruity with reference to contextual tone to obtain standardized incongruity, effectively implementing instance standardization for unbiased decision-making in MSD. ITFNet performs well in extracting salient and standardized incongruity through an incongruity-aware tension field, significantly tackling incongruity distortions and cross-instance variance. Moreover, ITFNet achieves state-of-the-art performance surpassing LLaVA1.5-7B with only 17.3M trainable parameters, demonstrating its optimal performance-efficiency in multi-modal sarcasm detection tasks.

Can Community Notes Replace Professional Fact-Checkers?

Nadav Borenstein, Greta Warren, Desmond Elliott, Isabelle Augenstein

Two commonly employed strategies to combat the rise of misinformation on social media are (i) fact-checking by professional organisations and (ii) community moderation by platform users. Policy changes by Twitter/X and, more recently, Meta, signal a shift away from partnerships with fact-checking organisations and towards an increased reliance on crowdsourced community notes. However, the extent and nature of dependencies between fact-checking and *helpful* community notes remain unclear. To address these questions, we use language models to annotate a large corpus of Twitter/X community notes with attributes such as topic, cited sources, and whether they refute claims tied to broader misinformation narratives. Our analysis reveals that community notes cite fact-checking sources up to five times more than previously reported. Fact-checking is especially crucial for notes on posts linked to broader narratives, which are *twice* as likely to reference fact-checking sources compared to other sources. Our results show that successful community moderation relies on professional fact-checking and highlight how citizen and professional fact-checking are deeply intertwined.

RecLM: Recommendation Instruction Tuning

Yangqin Jiang, Yuhao Yang, Lianghao Xia, Da Luo, Kangyi Lin, Chao Huang

Modern recommender systems aim to deeply understand users' complex preferences through their past interactions. While deep collaborative filtering approaches using Graph Neural Networks (GNNs) excel at capturing user-item relationships, their effectiveness is limited when handling sparse data or zero-shot scenarios, primarily due to constraints in ID-based embedding functions. To address these challenges, we propose a model-agnostic recommendation instruction-tuning paradigm that seamlessly integrates large language models with collaborative filtering. Our proposed Recommendation Language Model (RecLM) enhances the capture of user preference diversity through a carefully designed reinforcement learning reward function that facilitates self-augmentation of language models. Comprehensive evaluations demonstrate significant advantages of our approach across various settings, and its plug-and-play compatibility with state-of-the-art recommender systems results in notable performance enhancements.

Exploring LLMs' Ability to Spontaneously and Conditionally Modify Moral Expressions through Text Manipulation

Candida Maria Greco, Lucio La Cava, Lorenzo Zangari, Andrea Tagarelli

Morality serves as the foundation of societal structure, guiding legal systems, shaping cultural values, and influencing individual self-perception. With the rise and pervasiveness of generative AI tools, and particularly Large Language Models (LLMs), concerns arise regarding how these tools capture and potentially alter moral dimensions through machine-generated text manipulation. Based on the Moral Foundation Theory, our work investigates this topic by analyzing the behavior of 12 LLMs among the most widely used Open and uncensored (i.e., "abliterated") models, and leveraging human-annotated datasets used in moral-related analysis. Results have shown varying levels of alteration of moral expressions depending on the type of text modification task and moral-related conditioning prompt.

CULEMO: Cultural Lenses on Emotion - Benchmarking LLMs for Cross-Cultural Emotion Understanding

Tadesse Destaw Belay, Ahmed Haj Ahmed, Alvin C Grissom II, Iqra Ameer, Grigori Sidorov, Olga Kolesnikova, Seid Muhie Yimam

NLP research has increasingly focused on subjective tasks such as emotion analysis. However, existing emotion benchmarks suffer from two major shortcomings: (1) they largely rely on keyword-based emotion recognition, overlooking crucial cultural dimensions required for deeper emotion understanding, and (2) many are created by translating English-annotated data into other languages, leading to potentially unreliable evaluation. To address these issues, we introduce Cultural Lenses on Emotion (CuLEmo), the first benchmark designed to evaluate culture-aware emotion prediction across six languages: Amharic, Arabic, English, German, Hindi, and Spanish. CuLEmo comprises 400 crafted questions per language, each requiring nuanced cultural reasoning and understanding. We use this benchmark to evaluate several state-of-the-art LLMs on culture-aware emotion prediction and sentiment analysis tasks. Our findings reveal that (1) emotion conceptualizations vary significantly across languages and cultures, (2) LLMs performance likewise varies by language and cultural context, and (3) prompting in English with explicit country context often outperforms in-language prompts for culture-aware emotion and sentiment understanding. The dataset and evaluation code is available.

Language Model Fine-Tuning on Scaled Survey Data for Predicting Distributions of Public Opinions

Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, Serina Chang

Large language models (LLMs) present novel opportunities in public opinion research by predicting survey responses in advance during the early stages of survey design. Prior methods steer LLMs via descriptions of subpopulations as LLMs' input prompt, yet such prompt engineering approaches have struggled to faithfully predict the distribution of survey responses from human subjects. In this work, we propose directly fine-tuning LLMs to predict response distributions by leveraging unique structural characteristics of survey data. To enable fine-tuning, we curate SubPOP, a significantly scaled dataset of 3,362 questions and 70K subpopulation-response pairs from well-established public opinion surveys. We show that fine-tuning on SubPOP greatly improves the match between LLM predictions and human responses across various subpopulations, reducing the LLM-human gap by up to 46% compared to baselines, and achieves strong generalization to unseen surveys and subpopulations. Our findings highlight the potential of survey-based fine-tuning to improve opinion prediction for diverse, real-world subpopulations and therefore enable more efficient survey designs.

WhiSPA: Semantically and Psychologically Aligned Whisper with Self-Supervised Contrastive and Student-Teacher Learning

Rajath Rao, Adithya V Ganesan, Oscar Kjell, Jonah Luby, Akshay Raghavan, Scott M. Feltman, Whitney Ringwald, Ryan L. Boyd, Benjamin J. Luft, Camilo J. Ruggero, Neville Ryant, ROMAN KOTOV, H. Schwartz

Current speech encoding pipelines often rely on an additional text-based LM to get robust representations of human communication, even though SotA speech-to-text models often have a LM within. This work proposes an approach to improve the LM within an audio model such that the subsequent text-LM is unnecessary. We introduce **WhiSPA** (**Whi**sper with **S**emantic and **P**sychological **A**lignment), which leverages a novel audio training objective: contrastive loss with a language model embedding as a teacher. Using over 500k speech segments from mental health audio interviews, we evaluate the utility of aligning Whisper's latent space with semantic representations from a text autoencoder (SBERT) and lexically derived embeddings of basic psychological dimensions: emotion and personality. Over self-supervised affective tasks and downstream psychological tasks, WhiSPA surpasses current speech encoders, achieving an average error reduction of 73.4% and 83.8%, respectively. WhiSPA demonstrates that it is not always necessary to run a subsequent text LM on speech-to-text output in order to get a rich psychological representation of human communication.

Are We in the AI-Generated Text World Already? Quantifying and Monitoring AIGT on Social Media

Zhen Sun, Zongmin Zhang, Xinyue Shen, Ziyi Zhang, Yule Liu, Michael Backes, Yang Zhang, Xinlei He

Social media platforms are experiencing a growing presence of AI-Generated Texts (AIGTs). However, the misuse of AIGTs could have profound implications for public opinion, such as spreading misinformation and manipulating narratives. Despite its importance, it remains unclear how prevalent AIGTs are on social media. To address this gap, this paper aims to quantify and monitor the AIGTs on online social media platforms. We first collect a dataset (SM-D) with around 2.4M posts from 3 major social media platforms: Medium, Quora, and Reddit. Then, we construct a diverse dataset (AIGTBench) to train and evaluate AIGT detectors. AIGTBench combines popular open-source datasets and our AIGT datasets generated from social media texts by 12 LLMs, serving as a benchmark for evaluating mainstream detectors. With this setup, we identify the best-performing detector (OSM-Det). We then apply OSM-Det to SM-D to track AIGTs across social media platforms from January 2022 to October 2024, using the AI Attribution Rate (AAR) as the metric. Specifically, Medium and Quora exhibit marked increases in AAR, rising from 1.77\% to 37.03\% and 2.06\% to 38.95\%, respectively. In contrast, Reddit shows slower growth, with AAR increasing from 1.31\% to 2.45\% over the same period. Our further analysis indicates that AIGTs on social media differ from human-written texts across several dimensions, including linguistic patterns, topic distributions, engagement levels, and the follower distribution of authors. We envision our analysis and findings on AIGTs in social media can shed light on future research in this domain.

PCoT: Persuasion-Augmented Chain of Thought for Detecting Fake News and Social Media Disinformation

Arkadiusz Modzelewski, Witold Sosnowski, Tiziano Labruna, Adam Wierzbicki, Giovanni Da San Martino

Disinformation detection is a key aspect of media literacy. Psychological studies have shown that knowledge of persuasive fallacies helps individuals detect disinformation. Inspired by these findings, we experimented with large language models (LLMs) to test whether infusing persuasion knowledge enhances disinformation detection. As a result, we introduce the Persuasion-Augmented Chain of Thought (PCoT), a novel approach that leverages persuasion to improve disinformation detection in zero-shot classification. We extensively evaluate PCoT on online news and social media posts. Moreover, we publish two novel, up-to-date disinformation datasets: EUDisinfo and MultiDis. These datasets enable the evaluation of PCoT on content entirely unseen by the LLMs used in our experiments, as the content was published after the models' knowledge cutoffs. We show that, on average, PCoT outperforms competitive methods by 15\% across five LLMs and five datasets. These findings highlight the value of persuasion in strengthening zero-shot disinformation detection.

Mind the Gesture: Evaluating AI Sensitivity to Culturally Offensive Non-Verbal Gestures

Akhila Yerukola, Saadia Gabriel, Nanyun Peng, Maarten Sap

Gestures are an integral part of non-verbal communication, with meanings that vary across cultures, and misinterpretations that can have serious social and diplomatic consequences. As AI systems become more integrated into global applications, ensuring they do not inadvertently perpetuate cultural offenses is critical. To this end, we introduce Multi-Cultural Set of Inappropriate Gestures and Nonverbal Signs (MC-SIGNS), a dataset of 288 gesture-country pairs annotated for offensiveness, cultural significance, and contextual factors across 25 gestures and 85 countries. Through systematic evaluation using MC-SIGNS, we uncover critical limitations: text-to-image (T2I) systems exhibit strong US-centric biases, performing better at detecting offensive gestures in US contexts than in non-US ones; large language models (LLMs) tend to over-flag gestures as offensive; and vision-language models (VLMs) default to US-based interpretations when responding to universal concepts like wishing someone luck, frequently suggesting culturally inappropriate gestures. These findings highlight the urgent need for culturally-aware AI safety mechanisms to ensure equitable global deployment of AI technologies.

Mapping the Podcast Ecosystem with the Structured Podcast Research Corpus

Benjamin Roger Litterer, David Jurgens, Dallas Card

Podcasts provide highly diverse content to a massive listener base through a unique on-demand modality. However, limited data has prevented large-scale computational analysis of the podcast ecosystem. To fill this gap, we introduce a massive dataset of over 1.1M podcast transcripts that is largely comprehensive of all English language podcasts available through public RSS feeds from May and June of 2020. This data is not limited to text, but includes metadata, inferred speaker roles, and audio features and speaker turns for a subset of 370K episodes. Using this data, we conduct a foundational investigation into the content, structure, and responsiveness of this ecosystem. Together, our data and analyses open the door to continued computational research of this popular and impactful medium.

Meaning Variation and Data Quality in the Corpus of Founding Era American English

Dallas Card

Legal scholars are increasingly using corpus based methods for assessing historical meaning. Among work focused on the so-called founding era (mid to late 18th century), the majority of such studies use the Corpus of Founding Era American English (COFEA) and rely on methods such as word counting and manual coding. Here, we demonstrate what can be inferred about meaning change and variation using more advanced NLP methods, focusing on terms in the U.S. Constitution. We also carry out a data quality assessment of COFEA, pointing out issues with OCR quality and metadata, compare diachronic change to synchronic variation, and discuss limitations when using NLP methods for studying historical meaning.

Research Borderlands: Analysing Scientific Writing Across Research Cultures

Shaily Bhatt, Tal August, Maria Antoniak

Improving cultural competence of language technologies is important. However most recent works rarely engage with the communities they study, and instead rely on synthetic setups and imperfect proxies of culture. In this work, we take a human-centered approach to discover and measure language-based cultural norms, and cultural competence of LLMs. We focus on a single kind of culture, *research cultures*, and a single task, *adapting writing across research cultures*. Through a set of interviews with interdisciplinary researchers, who are experts at moving between cultures, we create a framework of structural, stylistic, rhetorical, and citational norms that vary across research cultures. We operationalise these features with a suite of computational metrics and use them for (a) surfacing latent cultural norms in human-written research papers at scale; and (b) highlighting the lack of cultural competence of LLMs, and their tendency to homogenize writing. Overall, our work illustrates the efficacy of a human-centered approach to measuring cultural norms in human-written and LLM-generated texts.

Counterspeech the ultimate shield! Multi-Conditioned Counterspeech Generation through Attributed Prefix Learning

Aswini Kumar Padhi, Anil Bandhakavi, Tanmoy Chakraborty

Counterspeech has proven to be a powerful tool to combat hate speech online. Previous studies have focused on generating counterspeech conditioned only on specific intents (single attributed). However, a holistic approach considering multiple attributes simultaneously can yield more nuanced and effective responses. Here, we introduce HiPPrO, Hierarchical Prefix learning with Preference Optimization, a novel two-stage framework that utilizes the effectiveness of attribute-specific prefix embedding spaces hierarchically optimized during the counterspeech generation process in the first phase. Thereafter, we incorporate both reference and reward-free preference optimization to generate more constructive counterspeech. Furthermore, we extend IntentCONANv2 by annotating all 13,973 counterspeech instances with emotion labels by five annotators. HiPPrO leverages hierarchical prefix optimization to integrate these dual attributes effectively. An extensive evaluation demonstrates that HiPPrO achieves a 38 % improvement in intent conformity and a 3 %, 2 %, 3 % improvement in Rouge-1, Rouge-2, and Rouge-L, respectively, compared to several baseline models. Human evaluations further substantiate the superiority of our approach, highlighting the enhanced relevance and appropriateness of the generated counterspeech. This work underscores the potential of multi-attribute conditioning in advancing the efficacy of counterspeech generation systems. Our code is available on Github and dataset is open-sourced on Hugging-face.

Beyond Negative Stereotypes – Non-Negative Abusive Utterances about Identity Groups and Their Semantic Variants

Tina Lommel, Elisabeth Eder, Josef Ruppenhofer, Michael Wiegand

We study a subtype of implicitly abusive language, namely non-negative sentences about identity groups (e.g. "Women make good cooks"), and introduce a novel dataset of such utterances. Not only do we profile such abusive sentences, but since our dataset includes different semantic variants of the same characteristic attributed to an identity group, we can also systematically study the impact of varying degrees of generalization and perspective framing. Similarly, we switch identity groups to assess whether the characteristic described in a sentence is inherently abusive. We also report on classification experiments.

Reviving Cultural Heritage: A Novel Approach for Comprehensive Historical Document Restoration

Yuyi Zhang, Peirong Zhang, Zhenhua Yang, Pengyu Yan, Yongxin Shi, Pengwei Liu, Fengjun Guo, Lianwen Jin

Historical documents represent an invaluable cultural heritage, yet have undergone significant degradation over time through tears, water erosion, and oxidation. Existing Historical Document Restoration (HDR) methods primarily focus on single modality or limited-size restoration, failing to meet practical needs. To fill this gap, we present a full-page HDR dataset (FPHDR) and a novel automated HDR solution (AutoHDR). Specifically, FPHDR comprises 1,633 real and 6,543 synthetic images with character-level and line-level locations, as well as character annotations in different damage grades. AutoHDR mimics historians' restoration workflows through a three-stage approach: OCR-assisted damage localization, vision-language context text prediction, and patch autoregressive appearance restoration. The modular architecture of AutoHDR enables seamless human-machine collaboration, allowing for flexible intervention and optimization at each restoration stage. Experiments demonstrate AutoHDR's remarkable performance in HDR. When processing severely damaged documents, our system improves OCR accuracy from 46.83% to 84.05%, with further enhancement to 94.25% through human-machine collaboration. We believe this work represents a significant advancement in automated historical document restoration and contributes substantially to cultural heritage preservation. The model and dataset are available at https://github.com/SCUT-DLVCLab/AutoHDR.

From Perceptions to Decisions: Wildfire Evacuation Decision Prediction with Behavioral Theory-informed LLMs

Ruxiao Chen, Chenguang Wang, Yuran Sun, Xilei Zhao, Susu Xu

Evacuation decision prediction is critical for efficient and effective wildfire response by helping emergency management anticipate traffic congestion and bottlenecks, allocate resources, and minimize negative impacts. Traditional statistical methods for evacuation decision prediction fail to capture the complex and diverse behavioral logic of different individuals. In this work, for the first time, we introduce *FLARE*, short for facilitating LLM for advanced reasoning on wildfire evacuation decision prediction, a Large Language Model (LLM)-based framework that integrates behavioral theories and models to streamline the Chain-of-Thought (CoT) reasoning and subsequently integrate with memory-based Reinforcement Learning (RL) module to provide accurate evacuation decision prediction and understanding. Our proposed method addresses the limitations of using existing LLMs for evacuation behavioral predictions, such as limited survey data, mismatching with behavioral theory, conflicting individual preferences, implicit and complex mental states, and intractable mental state-behavior mapping. Experiments on three post-wildfire survey datasets show an average of 20.47\% performance improvement over traditional theory-informed behavioral models, with strong cross-event generalizability. Our complete code is publicly available at https://github.com/SusuXu-s-Lab/FLARE

Hanging in the Balance: Pivotal Moments in Crisis Counseling Conversations

Vivian Nguyen, Lillian Lee, Cristian Danescu-Niculescu-Mizil

During a conversation, there can come certain moments where its outcome hangs in the balance. In these pivotal moments, how one responds can put the conversation on substantially different trajectories leading to significantly different outcomes. Systems that can detect when such moments arise could assist conversationalists in domains with highly consequential outcomes, such as mental health crisis counseling. In this work, we introduce an unsupervised computational method for detecting such pivotal moments as they happen. The intuition is that a moment is pivotal if our expectation of the conversation's outcome varies widely depending on what might be said next. By applying our method to crisis counseling conversations, we first validate it by showing that it aligns with human perception—counselors take significantly longer to respond during moments detected by our method—and with the eventual conversational trajectory—which is more likely to change course at these times. We then use our framework to explore the relation between the counselor's response during pivotal moments and the eventual outcome of the session.

-Stance: A Large-Scale Real World Dataset of Stances in Legal Argumentation

Ankita Gupta, Douglas Rice, Brendan O'Connor

We present -Stance, a large-scale dataset of stances involved in legal argumentation.-Stance contains stance-annotated argument pairs, semi-automatically mined from millions of examples of U.S. judges citing precedent in context using citation signals. The dataset aims to facilitate work on the legal argument stance classification task, which involves assessing whether a case summary strengthens or weakens a legal argument (polarity) and to what extent (intensity). To assess the complexity of this task, we evaluate various existing NLP methods, including zero-shot prompting proprietary large language models (LLMs), and supervised fine-tuning of smaller open-weight language models (LMs) on -Stance. Our findings reveal that although prompting proprietary LLMs can help predict stance polarity, supervised model fine-tuning on -Stance is necessary to distinguish intensity. We further find that alternative strategies such as domain-specific pretraining and zero-shot prompting using masked LMs remain insufficient. Beyond our dataset's utility for the legal domain, we further find that fine-tuning small LMs on -Stance improves their performance in other domains. Finally, we study how temporal changes in signal definition can impact model performance, highlighting the importance of careful data curation for downstream tasks by considering the historical and sociocultural context. We publish the associated dataset to foster further research on legal argument reasoning.

Only a Little to the Left: A Theory-grounded Measure of Political Bias in Large Language Models

Mats Faulborn, Indira Sen, Max Pellert, Andreas Spitz, David Garcia

Prompt-based language models like GPT4 and LLaMa have been used for a wide variety of use cases such as simulating agents, searching for information, or for content analysis. For all of these applications and others, political biases in these models can affect their performance. Several researchers have attempted to study political bias in language models using evaluation suites based on surveys, such as the Political Compass Test (PCT), often finding a particular leaning favored by these models. However, there is some variation in the exact prompting techniques, leading to diverging findings, and most research relies on constrained-answer settings to extract model responses. Moreover, the Political Compass Test is not a scientifically valid survey instrument. In this work, we contribute a political bias measured informed by political science theory, building on survey design principles to test a wide variety of input prompts, while taking into account prompt sensitivity. We then prompt 11 different open and commercial models, differentiating between instruction-tuned and non-instruction-tuned models, and automatically classify their political stances from 88,110 responses. Leveraging this dataset, we compute political bias profiles across different prompt variations and find that while PCT exaggerates bias in certain models like GPT3.5, measures of political bias are often unstable, but generally more left-leaning for instruction-tuned models. Code and data are available at https://github.com/MaFa211/theory_grounded_pol_bias.

The Noisy Path from Source to Citation: Measuring How Scholars Engage with Past Research

Hong Chen, Misha Teplitskiy, David Jurgens

Academic citations are widely used for evaluating research and tracing knowledge flows. Such uses typically rely on raw citation counts and neglect variability in citation types. In particular, citations can vary in their fidelity as original knowledge from cited studies may be paraphrased, summarized, or reinterpreted, possibly wrongly, leading to variation in how much information changes from cited to citing paper. In this study, we introduce a computational pipeline to quantify citation fidelity at scale. Using full texts of papers, the pipeline identifies citations in citing papers and the corresponding claims in cited papers, and applies supervised models to measure fidelity at the sentence level. Analyzing a large-scale multi-disciplinary dataset of approximately 13 million citation sentence pairs, we find that citation fidelity is higher when authors cite papers that are 1) more recent and intellectually close, 2) more accessible, and 3) the first author has a lower H-index and the author team is medium-sized. Using a quasi-experiment, we establish the "telephone effect" - when citing papers have low fidelity to the original claim, future papers that cite the citing paper and the original have lower fidelity to the original. Our work reveals systematic differences in citation fidelity, underscoring the limitations of analyses that rely on citation quantity alone and the potential for distortion of evidence.

Towards Style Alignment in Cross-Cultural Translation

Shreya Havaldar, Adam Stein, Eric Wong, Lyle Ungar

Successful communication depends on the speaker's intended style (i.e., what the speaker is trying to convey) aligning with the listener's interpreted style (i.e., what the listener perceives). However, cultural differences often lead to misalignment between the two; for example, politeness is often lost in translation. We characterize the ways that LLMs fail to translate style -- biasing translations towards neutrality and performing worse in non-Western languages. We mitigate these failures with RASTA (Retrieval-Augmented STylistic Alignment), a method that leverages learned stylistic concepts to encourage LLM translation to appropriately convey cultural communication norms and align style.

A Multi-Agent Framework for Mitigating Dialect Biases in Privacy Policy Question-Answering Systems

Đorđe Klisura, Astrid R Bernaga Torres, Anna Karen Gárate-Escamilla, Rajesh Roshan Biswal, Ke Yang, Hilal Pataci, Anthony Rios

Privacy policies inform users about data collection and usage, yet their complexity limits accessibility for diverse populations. Existing Privacy Policy Question Answering (QA) systems exhibit performance disparities across English dialects, disadvantaging speakers of non-standard varieties. We propose a novel multi-agent framework inspired by human-centered design principles to mitigate dialectal biases. Our approach integrates a Dialect Agent, which translates queries into Standard American English (SAE) while preserving dialectal intent, and a Privacy Policy Agent, which refines predictions using domain expertise. Unlike prior approaches, our method does not require retraining or dialect-specific fine-tuning, making it broadly applicable across models and domains. Evaluated on PrivacyQA and PolicyQA, our framework improves GPT-4o-mini's zero-shot accuracy from 0.394 to 0.601 on PrivacyQA and from 0.352 to 0.464 on PolicyQA, surpassing or matching few-shot baselines without additional training data. These results highlight the effectiveness of structured agent collaboration in mitigating dialect biases and underscore the importance of designing NLP systems that account for linguistic diversity to ensure equitable access to privacy information.

SocialDuolingo: Interactive Evaluation for Cultural Competence in Language Agents

Jincenzi Wu, Jianxun Lian, Dingdong WANG, Helen M. Meng

Large Language Models (LLMs) are increasingly deployed worldwide, yet their ability to navigate cultural nuances remains underexplored. Misinterpreting cultural content can lead to AI-generated responses that are offensive or inappropriate, limiting their usability in global applications such as customer service, diplomatic communication, and online education. While prior research has evaluated cultural knowledge of LLMs, existing benchmarks fail to assess dynamic cultural competence-the ability to apply cultural knowledge effectively in real-world interactions. To address this gap, we introduce SocialDuolingo, a novel benchmark designed to evaluate cultural competence through multi-turn interactive intercultural scenarios. It comprises 3,060 human-written scenarios spanning 60 countries across six continents. Through extensive experiments on eight prominent LLMs, our findings reveal a significant gap between the cultural knowledge stored in these models and their ability to apply it effectively in cross-cultural communication.

Session 12: Discourse and Pragmatics (Wednesday 30 July, 11:00-12:30, Hall 4/5)

Towards LLM-powered Attentive Listener: A Pragmatic Approach through Quantity Self-Repair

Junlin Li, Bo Peng, Yu-Yin Hsu

Grice's Quantity Maxims dictate that human speakers aim for the optimal quantity of information during conversation. To empower LLMs to self-repair their responses toward optimal quantity and improve their attentive listening skills, we propose Q-Tuning and Q-Traveling, which draw on heuristic path-finding to enable decoder-only LLMs to travel among multiple ``Q-alternatives" (Quantity Alternatives) and search for the optimal quantity in coordination with a conversation goal. Automatic and human evaluations demonstrate the effectiveness of Q-Tuning and Q-Traveling in constructing human-like, user-centered conversation agents.

No Questions are Stupid, but some are Poorly Posed: Understanding Poorly-Posed Information-Seeking Questions

Neha Srikanth, Rachel Rudinger, Jordan Lee Boyd-Graber

Questions help unlock information to satisfy users' information needs. However, when the question is poorly posed, answerers (whether human or computer) may struggle to answer the question in a way that satisfies the asker, despite possibly knowing everything necessary to address the asker's latent information need. Using Reddit question-answer interactions from r/NoStupidQuestions, we develop a computational framework grounded in linguistic theory to study poorly-posedness of questions by generating spaces of potential interpretations of questions and computing distributions over these spaces based on interpretations chosen by both human answerers in the Reddit question thread, as well as by a suite of large language models. Both humans and models struggle to converge on dominant interpretations when faced with poorly-posed questions, but employ different strategies: humans focus on specific interpretations through question negotiation, while models attempt comprehensive coverage by addressing many interpretations simultaneously.

Discourse Relation-Enhanced Neural Coherence Modeling

Wei Liu, Michael Strube

Discourse coherence theories posit relations between text spans as a key feature of coherent texts. However, existing work on coherence modeling has paid little attention to discourse relations. In this paper, we provide empirical evidence to demonstrate that relation features are correlated with text coherence. Then, we investigate a novel fusion model that uses position-aware attention and a visible matrix to combine text- and relation-based features for coherence assessment. Experimental results on two benchmarks show that our approaches can significantly improve baselines, demonstrating the importance of relation features for coherence modeling.

Probing LLMs for Multilingual Discourse Generalization Through a Unified Label Set

Florian Eichin, Yang Janet Liu, Barbara Plank, Michael A. Hedderich

Discourse understanding is essential for many NLP tasks, yet most existing work remains constrained by framework-dependent discourse representations. This work investigates whether large language models (LLMs) capture discourse knowledge that generalizes across languages and frameworks. We address this question along two dimensions: (1) developing a unified discourse relation label set to facilitate cross-lingual and cross-framework discourse analysis, and (2) probing LLMs to assess whether they encode generalizable discourse abstractions. Using multilingual discourse relation classification as a testbed, we examine a comprehensive set of 23 LLMs of varying sizes and multilingual capabilities. Our results show that LLMs, especially those with multilingual training corpora, can generalize discourse information across languages and frameworks. Further layer-wise analyses reveal that language generalization at the discourse level is most salient in the intermediate layers. Lastly, our error analysis provides an account of challenging relation classes.

Improving Dialogue Discourse Parsing through Discourse-aware Utterance Clarification

Yaxin Fan, Peifeng Li, Qiaoming Zhu

Dialogue discourse parsing aims to identify and analyze discourse relations between the utterances within dialogues. However, linguistic features in dialogues, such as omission and idiom, frequently introduce ambiguities that obscure the intended discourse relations, posing significant challenges for parsers. To address this issue, we propose a Discourse-aware Clarification Module (DCM) to enhance the performance of the dialogue discourse parser. DCM employs two distinct reasoning processes: clarification type reasoning and discourse goal reasoning. The former analyzes linguistic features, while the latter distinguishes the intended relation from the ambiguous one. Furthermore, we introduce Contribution-aware Preference Optimization (CPO) to mitigate the risk of erroneous clarifications, thereby reducing cascading errors. CPO enables the parser to assess the contributions of the clarifications from DCM and provide feedback to optimize the DCM, enhancing its adaptability and alignment with the parser's requirements. Extensive experiments on the STAC and Molweni datasets demonstrate that our approach effectively resolves ambiguities and significantly outperforms the state-of-the-art (SOTA) baselines.

(RSA)2: A Rhetorical-Strategy-Aware Rational Speech Act Framework for Figurative Language Understanding

Cesare Spinoso-Di Piano, David Eric Austin, Pablo Piantanida, Jackie CK Cheung

Figurative language (e.g., irony, hyperbole, understatement) is ubiquitous in human communication, resulting in utterances where the literal and the intended meanings do not match. The Rational Speech Act (RSA) framework, which explicitly models speaker intentions, is the most widespread theory of probabilistic pragmatics, but existing implementations are either unable to account for figurative expressions or require modeling the implicit motivations for using figurative language (e.g., to express joy or annoyance) in a setting-specific way. In this paper, we introduce the Rhetorical-Strategy-Aware RSA (RSA)^2 framework which models figurative language use by considering a speaker's employed rhetorical strategy. We show that (RSA)^2 enables human-compatible interpretations of non-literal utterances without modeling a speaker's motivations for being non-literal. Combined with LLMs, it achieves state-of-the-art performance on the ironic split of PragMega+, a new irony interpretation dataset introduced in this study.

BOOKCOREF: Coreference Resolution at Book Scale

Giuliano Martinelli, Tommaso Bonomo, Pere-Lluís Huguet Cabot, Roberto Navigli

Coreference Resolution systems are typically evaluated on benchmarks containing small- to medium-scale documents. When it comes to evaluating long texts, however, existing benchmarks, such as LitBank, remain limited in length and do not adequately assess system capabilities at the book scale, i.e., when co-referring mentions span hundreds of thousands of tokens. To fill this gap, we first put forward a novel automatic pipeline that produces high-quality Coreference Resolution annotations on full narrative texts. Then, we adopt this pipeline to create the first book-scale coreference benchmark, BOOKCOREF, with an average document length of more than 200,000 tokens. We carry out a series of experiments showing the robustness of our automatic procedure and demonstrating the value of our resource, which enables current long-document coreference systems to gain up to +20 CoNLL-F1 points when evaluated on full books. Moreover, we report on the new challenges introduced by this unprecedented book-scale setting, highlighting that current models fail to deliver the same performance they achieve on smaller documents. We release our data and code to encourage research and development of new book-scale Coreference Resolution systems at https://github.com/sapienzanlp/bookcoref.

Session 12: Interpretability and Analysis of Models for NLP (Wednesday 30 July, 11:00-12:30, Hall 4/5)

Can Input Attributions Interpret the Inductive Reasoning Process in In-Context Learning?

Mengyu Ye, Tatsuki Kuribayashi, Goro Kobayashi, Jun Suzuki

Interpreting the internal process of neural models has long been a challenge. This challenge remains relevant in the era of large language models (LLMs) and in-context learning (ICL); for example, ICL poses a new issue of interpreting which example in the few-shot examples contributed to identifying/solving the task. To this end, in this paper, we design synthetic diagnostic tasks of inductive reasoning, inspired by the generalization tests in linguistics; here, most in-context examples are ambiguous w.r.t. their underlying rule, and one critical example disambiguates the task demonstrated. The question is whether conventional input attribution (IA) methods can track such a reasoning process, i.e., identify the influential example, in ICL. Our experiments provide several practical findings; for example, a certain simple IA method works the best, and the larger the model, the generally harder it is to interpret the ICL with gradient-based IA methods.

The Reasoning-Memorization Interplay in Language Models Is Mediated by a Single Direction

Yihuai Hong, Meng Cao, Dian Zhou, Lei Yu, Zhijing Jin

Large language models (LLMs) excel on a variety of reasoning benchmarks, but previous studies suggest they sometimes struggle to generalize to unseen questions, potentially due to over-reliance on memorized training examples. However, the precise conditions under which LLMs switch between reasoning and memorization during text generation remain unclear. In this work, we provide a mechanistic understanding of LLMs' reasoning-memorization dynamics by identifying a set of linear features in the model's residual stream that govern the balance between genuine reasoning and memory recall. These features not only distinguish reasoning tasks from memory-intensive ones but can also be manipulated to causally influence model performance on reasoning tasks. Additionally, we show that intervening in these reasoning features helps the model more accurately activate the most relevant problem-solving capabilities during answer generation. Our findings offer new insights into the underlying mechanisms of reasoning and memory in LLMs and pave the way for the development of more robust and interpretable generative AI systems.

Explicit Bayesian Inference to Uncover the Latent Themes of Large Language Models

Raymond Li, Chuyuan Li, Gabriel Murray, Giuseppe Carenini

Large language models (LLMs) have demonstrated impressive generative capabilities, yet their inner mechanisms remain largely opaque. In this work, we introduce a novel approach to interpret LLMs generation process through the lens of an explicit Bayesian framework by inferring latent topic variables via variational inference. Specifically, we leverage a variational autoencoder-based neural topic model to dynamically approximate the posterior distribution over the high-level latent topic variables at each generation step. By reconstructing the LLM's next-token predictions through these latent topics and maintaining a regularized latent space, our method yields interpretable and diverse topic representations but also has the ability to effectively captures semantic shifts throughout the text. We validate our approach on multiple datasets, showing that our latent topics outperform state-of-the-art topic models on intrinsic measures of coherence and diversity. Furthermore, we demonstrate the utility of our approach in downstream applications by using the inferred topic distributions to retrieve relevant demonstration examples for in-context learning, resulting in significant gains on classification and summarization tasks.

DPO Kernels: A Semantically-Aware, Kernel-Enhanced, and Divergence-Rich Paradigm for Direct Preference Optimization

Amitava Das, Suranjana Trivedy, Danush Khanna, Yaswanth Narsupalli, Basab Ghosh, Rajarshi Roy, Gurpreet Singh, Vinija Jain, Vasu Sharma, Aishwarya Naresh Reganti, Aman Chadha

The rapid advancement of large language models (LLMs) has revolutionized numerous applications, but presents significant challenges in aligning these models with diverse human values, ethical standards, and specific user preferences. Direct Preference Optimization (DPO) has become a cornerstone for preference alignment but is constrained by reliance on fixed divergence measures and limited feature transformations. We introduce \textbf{DPO-Kernels}, an innovative enhancement of DPO that integrates kernel methods to overcome these challenges through four key contributions: (i) \textbf{Kernelized Representations}: These representations enhance divergence measures by using polynomial, RBF, Mahalanobis, and spectral kernels for richer feature transformations. Additionally, we introduce a \textbf{hybrid loss} that combines embedding-based loss with probability-based loss; (ii) \textbf{Divergence Alternatives}: Beyond Kullback–Leibler (KL), we incorporate Jensen-Shannon, Hellinger, Rényi, Bhattacharyya, Wasserstein, and other f-divergences to boost stability and robustness; (iii) \textbf{Data-Driven Selection}: Choosing the optimal kernel-divergence pair among 28 combinations (4 kernels \times 7 divergences) is challenging. We introduce automatic metrics that analyze the data to select the best kernel-divergence pair, eliminating the need for manual tuning; (iv) \textbf{Hierarchical Mixture of Kernels (HMK)}: Combining local and global kernels for precise and large-scale semantic modeling. This approach automatically selects the optimal kernel mixture during training, enhancing modeling flexibility. DPO-Kernels achieve state-of-the-art generalization in factuality, safety, reasoning, and instruction following across 12 datasets. While alignment risks overfitting, Heavy-Tailed Self-Regularization (HT-SR) theory confirms that DPO-Kernels ensure robust generalization in LLMs. Comprehensive resources are \href{https://github.com/anonymous-panda123/DPO-Kernels}{available} to facilitate further research and application of DPO-Kernels.

Lost in Transcription, Found in Distribution Shift: Demystifying Hallucination in Speech Foundation Models

Hanin atwany, Abdul Waheed, Rita Singh, Monojit Choudhury, Bhiksha Raj

Speech foundation models trained at a massive scale, both in terms of model and data size, result in robust systems capable of performing multiple speech tasks, including automatic speech recognition (ASR). These models transcend language and domain barriers, yet effectively measuring their performance remains a challenge. Traditional metrics like word error rate (WER) and character error rate (CER) are commonly used to evaluate ASR performance but often fail to reflect transcription quality in critical contexts, particularly when detecting fabricated outputs. This phenomenon, known as hallucination, is especially concerning in high-stakes domains such as healthcare, legal, and aviation, where errors can have severe consequences. In our work, we address this gap by investigating hallucination in ASR models. We examine how factors such as distribution shifts, model size, and model architecture influence the hallucination error rate (HER), a metric we introduce to quantify hallucinations. Our analysis of 20 ASR models reveals 3 key insights: (1) High WERs can mask low hallucination rates, while low WERs may conceal dangerous hallucinations. (2) Synthetic noise, both adversarial and common perturbations like white noise, pitch shift, and time stretching, increase HER. (3) Distribution shift correlates strongly with HER ((= 0.91)). Our findings highlight the importance of incorporating HER alongside traditional metrics like WER to better assess ASR model performance, particularly in high-stakes domains.

The Law of Knowledge Overshadowing: Towards Understanding, Predicting and Preventing LLM Hallucination

Yuji Zhang, Sha Li, Cheng Qian, Jiateng Liu, Pengfei Yu, Chi Han, Yi R. Fung, Kathleen McKeown, ChengXiang Zhai, Manling Li, Heng Ji

Hallucination is a persistent challenge in large language models (LLMs), where even with rigorous quality control, models often generate distorted facts. This paradox, in which error generation continues despite high-quality training data, calls for a deeper understanding of the underlying LLM mechanisms. To address it, we propose a novel concept: knowledge overshadowing, where model's dominant knowledge can obscure less prominent knowledge during text generation, causing the model to fabricate inaccurate details. Building on this idea, we introduce a novel framework to quantify factual hallucinations by modeling knowledge overshadowing. Central to our approach is the log-linear law, which predicts that the rate of factual hallucination increases linearly with the logarithmic scale of (1) Knowledge Popularity, (2) Knowledge Length, and (3) Model Size. The law provides a means to preemptively quantify hallucinations, offering foresight into their occurrence even before model training or inference. Built on overshadowing effect, we propose a new decoding strategy CoDa, to mitigate hallucinations, which notably enhance model factuality on Overshadow (27.9%), MemoTrap (13.1%) and NQ-Swap (18.3%). Our findings not only deepen understandings of the underlying mechanisms behind hallucinations but also provide actionable insights for developing more predictable and controllable language models.

Un-considering Contextual Information: Assessing LLMs' Understanding of Indexical Elements

Metehan Oğuz, Yavuz Faruk Bakman, Duygu Nur Yaldiz

Large Language Models (LLMs) have demonstrated impressive performances in tasks related to coreference resolution. However, previous studies mostly assessed LLM performance on coreference resolution with nouns and third person pronouns. This study evaluates LLM performance on coreference resolution with indexical like I, you, here and tomorrow} which come with unique challenges due to their linguistic properties. We present the first study examining how LLMs interpret indexicals in English, releasing the English Indexical Dataset with 1600 multiple-choice questions. We evaluate pioneering LLMs, including GPT-4o, Claude 3.5 Sonnet, Gemini 1.5 Pro, and DeepSeek V3. Our results reveal that LLMs exhibit an impressive performance with some indexicals (I), while struggling with others (you, here, tomorrow), and that syntactic cues (e.g. quotation) contribute to LLM performance with some indexicals, while they reduce performance with others.

Time Travel: A Comprehensive Benchmark to Evaluate LMMs on Historical and Cultural Artifacts

Sara Ghaboura, Ketan Pravin More, Ritesh Thawkar, Wafa Al Ghallabi, Omkar Thawakar, Fahad Shahbaz Khan, Hisham Cholakkal, Salman Khan, Rao Muhammad Anwer

Understanding historical and cultural artifacts demands human expertise and advanced computational techniques, yet the process remains complex and time-intensive. While large multimodal models offer promising support, their evaluation and improvement require a standardized benchmark. To address this, we introduce TimeTravel, a benchmark of 10,250 expert-verified samples spanning 266 distinct cultures across 10 major historical regions. Designed for AI-driven analysis of manuscripts, artworks, inscriptions, and archaeological discoveries, TimeTravel provides a structured dataset and robust evaluation framework to assess AI models' capabilities in classification, interpretation, and historical comprehension. By integrating AI with historical research, TimeTravel fosters AI-powered tools for historians, archaeologists, researchers, and cultural tourists to extract valuable insights while ensuring technology contributes meaningfully to historical discovery and cultural heritage preservation. We evaluate contemporary AI models on TimeTravel, highlighting their strengths and identifying areas for improvement. Our goal is to establish AI as a reliable partner in preserving cultural heritage, ensuring that technological advancements contribute meaningfully to historical discovery.

Unveiling and Addressing Pseudo Forgetting in Large Language Models

Huashan Sun, Yizhe Yang, Yinghao Li, Jiawei Li, Yang Gao

Although substantial efforts have been made to mitigate catastrophic forgetting in continual learning, the intrinsic mechanisms are not well understood. In this work, we demonstrate the existence of "pseudo forgetting": the performance degradation in previous tasks is not attributed to a loss of capabilities, but rather to the failure of the instructions to activate the appropriate model capabilities. We show that the model's performance on previous tasks can be restored through two simple interventions: (1) providing partial external correct rationale, and (2) appending semantically meaningless suffixes to the original instructions, to guide the generation of correct rationales. Through empirical analysis of the internal mechanisms governing rationale generation, we reveal that models exhibiting pseudo forgetting show reduced instruction dependence during rationale generation, leading to suboptimal activation of their inherent capabilities. Based on this insight, we propose Rationale-Guidance Difficulty based Replay (RGD-R) framework that dynamically allocates replay data based on the model's ability to correctly leverage the intrinsic capabilities. Experimental results demonstrate that RGD-R effectively mitigates pseudo forgetting while maintaining model plasticity.

COVER: Context-Driven Over-Refusal Verification in LLMs

Giovanni Sullutrone, Riccardo A. Vigliermo, Sonia Bergamaschi, Luca Sala

We introduce the concept of context-driven over-refusal, an abstention arising when model's safety guardrails are triggered by the grounding knowledge provided alongside the user's request. Distinct from question-driven over-refusal, this occurs in both retrieval-augmented generation (RAG) and natural language processing (NLP) task completion (e.g. summarization, translation) where external content can unexpectedly trigger refusals. In this work, we present a novel two-stage evaluation framework named COVER, designed to quantify and analyze this behavior. Through a comprehensive empirical study on two public corpora, we show that over-refusal rates strongly depend on the task, system prompts, model family, and the number of retrieved documents. We observe that tasks such as translation and summarization yield disproportionately high over-refusal rates, while question-answering remains relatively robust, especially in newer models. Moreover, increasing the number of contextual documents tends to reduce refusals, yet broadens the pool of prompts at risk of encountering at least one "unsafe" text. Interestingly, strict system prompts do not necessarily lead to higher over-refusal rates, suggesting that in the absence of explicit directives, some models may default to a more cautious behavior. These findings highlight the need for fine-grained alignment and benchmarking strategies sensitive to both user intent and contextual nuances, offering a roadmap for future research in model training and evaluation.

COSMIC: Generalized Refusal Identification in LLM Activations

Vincent Siu, Nicholas Crispino, Zihao Yu, Sam Pan, Zhun Wang, Yang Liu, Dawn Song, Chenguang Wang

Large Language Models encode behaviors like refusal within their activation space, but identifying these behaviors remains challenging. Existing methods depend on predefined refusal templates detectable in output tokens or manual review. We introduce **COSMIC** (Cosine Similarity Metrics for Inversion of Concepts), an automated framework for direction selection that optimally identifies steering directions and target layers using cosine similarity, entirely independent of output text. COSMIC achieves steering effectiveness comparable to prior work without any prior knowledge or assumptions of a model's refusal behavior such as the use of certain refusal tokens. Additionally, COSMIC successfully identifies refusal directions in adversarial scenarios and models with weak safety alignment, demonstrating its robustness across diverse settings.

Do Language Models Mirror Human Confidence? Exploring Psychological Insights to Address Overconfidence in LLMs

Chenjun Xu, Bingbing Wen, Bin HAN, Robert Wolfe, Lucy Lu Wang, Bill Howe

Psychology research has shown that humans are poor at estimating their performance on tasks, tending towards underconfidence on easy tasks and overconfidence on difficult tasks. We examine three LLMs, Llama-3-70B-instruct, Claude-3-Sonnet, and GPT-4o, on a range of QA tasks of varying difficulty, and show that models exhibit subtle differences from human patterns of overconfidence: less sensitive to task difficulty, and when prompted to answer based on different personas---e.g., expert vs layman, or different race, gender, and ages---the models will respond with stereotypically biased confidence estimations even though their underlying answer accuracy remains the same. Based on these observations, we propose a novel method, Answer-Free Confidence Estimation (AFCE), for improving confidence calibration and LLM interpretability in these settings. AFCE is a self-assessment method that employs two stages of prompting, first eliciting only confidence scores on questions without answers, then asking separately for the answer. Experiments on MMLU and GPQA datasets spanning subjects and difficulty show that AFCE significantly improves confidence calibration performance for the most difficult problems where overconfidence is most pronounced.

Spotting Out-of-Character Behavior: Atomic-Level Evaluation of Persona Fidelity in Open-Ended Generation

Jisu Shin, Juhyun Oh, Eunsu Kim, Hoyun Song, Alice Oh

Ensuring persona fidelity in large language models (LLMs) is essential for maintaining coherent and engaging human-AI interactions. However, LLMs often exhibit Out-of-Character (OOC) behavior, where generated responses deviate from an assigned persona, leading to inconsistencies that affect model reliability. Existing evaluation methods typically assign single scores to entire responses, struggling to capture subtle persona misalignment, particularly in long-form text generation. To address this limitation, we propose an atomic-level evaluation framework that quantifies persona fidelity at a finer granularity. Our three key metrics measure the degree of persona alignment and consistency within and across generations. Our approach enables a more precise and realistic assessment of persona fidelity by identifying subtle deviations that real users would encounter. Through our experiments, we demonstrate that our framework effectively detects persona inconsistencies that prior methods overlook. By analyzing persona fidelity across diverse tasks and personality types, we reveal how task structure and persona desirability influence model adaptability, highlighting challenges in maintaining consistent persona expression.

RomanLens: The Role Of Latent Romanization In Multilinguality In LLMs

Alan Saji, Jaavid Aktar Husain, Thanmay Jayakumar, Raj Dabre, Anoop Kunchukuttan, Ratish Puduppully

Large Language Models (LLMs) exhibit remarkable multilingual generalization despite being predominantly trained on English-centric corpora. A fundamental question arises: how do LLMs achieve such robust multilingual capabilities? We take the case of non-Roman script languages, we investigate the role of Romanization—the representation of non-Roman scripts using Roman characters—as a bridge in multilingual processing. Using mechanistic interpretability techniques, we analyze next-token generation and find that intermediate layers frequently represent target words in Romanized form before transitioning to native script, a phenomenon we term Latent Romanization. Further, through activation patching experiments, we demonstrate that LLMs encode semantic concepts similarly across native and Romanized scripts, suggesting a shared underlying representation. Additionally, for translation into non-Roman script languages, our findings reveal that when the target language is in Romanized form, its representations emerge earlier in the model's layers compared to native script. These insights contribute to a deeper understanding of multilingual representation in LLMs and highlight the implicit role of Romanization in facilitating language transfer.

Steering into New Embedding Spaces: Analyzing Cross-Lingual Alignment Induced by Model Interventions in Multilingual Language Models

Anirudh Sundar, Sinead Williamson, Katherine Metcalf, Barry-John Theobald, Skyler Seto, Masha Fedzechkina

Aligned representations across languages is a desired property in multilingual large language models (mLLMs), as alignment can improve performance in cross-lingual tasks. Typically alignment requires fine-tuning a model, which is computationally expensive, and sizable language data, which often may not be available. A data-efficient alternative to fine-tuning is model interventions --- a method for manipulating model activations to steer generation into the desired direction. We analyze the effect of a popular intervention (finding experts) on the alignment of cross-lingual representations in mLLMs. We identify the neurons to manipulate for a given language and introspect the embedding space of mLLMs pre- and post-manipulation. We show that modifying the mLLM's activations changes its embedding space such that cross-lingual alignment is enhanced. Further, we show that the changes to the embedding space translate into improved downstream performance on retrieval tasks, with up to 2x improvements in top-1 accuracy on cross-lingual retrieval.

Unveiling Language-Specific Features in Large Language Models via Sparse Autoencoders

Boyi Deng, Yu Wan, Baosong Yang, Yidan Zhang, Fuli Feng

The mechanisms behind multilingual capabilities in Large Language Models (LLMs) have been examined using neuron-based or internal-activation-based methods. However, these methods often face challenges such as superposition and layer-wise activation variance, which limit their reliability. Sparse Autoencoders (SAEs) offer a more nuanced analysis by decomposing the activations of LLMs into a sparse linear combination of SAE features. We introduce a novel metric to assess the monolinguality of features obtained from SAEs, discovering that some features are strongly related to specific languages. Additionally, we show that ablating these SAE features only significantly reduces abilities in one language of LLMs, leaving others almost unaffected. Interestingly, we find some languages have multiple synergistic SAE features, and ablating them together yields greater improvement than ablating individually. Moreover, we leverage these SAE-derived language-specific features to enhance steering vectors, achieving control over the language generated by LLMs. The code is publicly available at.

Stepwise Reasoning Disruption Attack of LLMs

Jingyu Peng, Maolin Wang, Xiangyu Zhao, Kai Zhang, Wanyu Wang, Pengyue Jia, Qidong Liu, Ruocheng Guo, Qi Liu

Large language models (LLMs) have made remarkable strides in complex reasoning tasks, but their safety and robustness in reasoning processes remain unexplored, particularly in third-party platforms that facilitate user interactions via APIs. Existing attacks on LLM reasoning are constrained by specific settings or lack of imperceptibility, limiting their feasibility and generalizability. To address these challenges, we propose the Stepwise rEasoning Error Disruption (SEED) attack, which subtly injects errors into prior reasoning steps to mislead the model into producing incorrect subsequent reasoning and final answers. Unlike previous methods, SEED is compatible with zero-shot and few-shot settings, maintains the natural reasoning flow, and ensures covert execution without modifying the instruction. Extensive experiments on four datasets across four different models demonstrate SEED's effectiveness, revealing the vulnerabilities of LLMs to disruptions in reasoning processes. These findings underscore the need for greater attention to the robustness of LLM reasoning to ensure safety in practical applications. Our code is available at: https://github.com/Applied-Machine-Learning-Lab/SEED-Attack

Lost in Multilinguality: Dissecting Cross-lingual Factual Inconsistency in Transformer Language Models

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, Hinrich Schuetze

Multilingual language models (MLMs) store factual knowledge across languages but often struggle to provide consistent responses to semantically equivalent prompts in different languages. While previous studies point out this cross-lingual inconsistency issue, the underlying causes remain unexplored. In this work, we use mechanistic interpretability methods to investigate cross-lingual inconsistencies in MLMs. We find that MLMs encode knowledge in a language-independent concept space through most layers, and only transition to language-specific spaces in the final layers. Failures during the language transition often result in incorrect predictions in the target language, even when the answers are correct in other languages. To mitigate this inconsistency issue, we propose a linear shortcut method that bypasses computations in the final layers, enhancing both prediction accuracy and cross-lingual consistency. Our findings shed light on the internal mechanisms of MLMs and provide a lightweight, effective strategy for producing more consistent factual outputs.

The TIP of the Iceberg: Revealing a Hidden Class of Task-in-Prompt Adversarial Attacks on LLMs

Sergey Berezin, Reza Farahbakhsh, Noel Crespi

We present a novel class of jailbreak adversarial attacks on LLMs, termed Task-in-Prompt (TIP) attacks. Our approach embeds sequence-to-sequence tasks (e.g., cipher decoding, riddles, code execution) into the model’s prompt to indirectly generate prohibited inputs. To systematically assess the effectiveness of these attacks, we introduce the PHRYGE benchmark. We demonstrate that our techniques successfully circumvent safeguards in six state-of-the-art language models, including GPT-4o and LLaMA 3.2. Our findings highlight critical weaknesses in current LLM safety alignment and underscore the urgent need for more sophisticated defence strategies.

VLM2-Bench: A Closer Look at How Well VLMs Implicitly Link Explicit Matching Visual Cues

Jianshu Zhang, Dongyu Yao, Renjie Pi, Paul Pu Liang, Yi R. Fung

Visually linking matching cues is a crucial ability in daily life, such as identifying the same person in multiple photos based on their cues, even without knowing who they are. Despite the extensive knowledge that vision-language models (VLMs) possess, it remains largely unexplored whether they are capable of performing this fundamental task. To address this, we introduce VLM2-Bench, a benchmark designed to assess whether VLMs can Visually Link Matching cues, with 9 subtasks and over 3,000 test cases. Comprehensive evaluation across twelve VLMs, along with further analysis of various language-side and vision-side prompting methods, leads to a total of eight key findings. We identify critical challenges in models' ability to link visual cues, highlighting a significant performance gap. Based on these insights, we advocate for (i) enhancing core visual capabilities to improve adaptability and reduce reliance on prior knowledge, (ii) establishing clearer principles for integrating language-based reasoning in vision-centric tasks to prevent unnecessary biases, and (iii) shifting vision-text training paradigms toward fostering models' ability to independently structure and infer relationships among visual cues.

Enough Coin Flips Can Make LLMs Act Bayesian

Ritwik Gupta, Rodolfo Corona, Jiaxin Ge, Eric Wang, Dan Klein, Trevor Darrell, David M. Chan

Large language models (LLMs) exhibit the ability to generalize given few-shot examples in their input prompt, an emergent capability known as in-context learning (ICL). We investigate whether LLMs use ICL to perform structured reasoning in ways that are consistent with a Bayesian framework or rely on pattern matching. Using a controlled setting of biased coin flips, we find that: (1) LLMs often possess biased priors, causing initial divergence in zero-shot settings, (2) in-context evidence outweighs explicit bias instructions, (3) LLMs broadly follow Bayesian posterior updates, with deviations primarily due to miscalibrated priors rather than flawed updates, and (4) attention magnitude has negligible effect on Bayesian inference. With sufficient demonstrations of biased coin flips via ICL, LLMs update their priors in a Bayesian manner. Code and visualizations are available on the [project page](https://ai-climate.berkeley.edu/llm-coin-flips/).

Collapse of Dense Retrievers: Short, Early, and Literal Biases Outranking Factual Evidence

Mohsen Fayyaz, Ali Modarressi, Hinrich Schuetze, Nanyun Peng

Dense retrieval models are commonly used in Information Retrieval (IR) applications, such as Retrieval-Augmented Generation (RAG). Since they often serve as the first step in these systems, their robustness is critical to avoid downstream failures. In this work, we repurpose a relation extraction dataset (e.g., Re-DocRED) to design controlled experiments that quantify the impact of heuristic biases, such as a preference for shorter documents, on retrievers like Dragon+ and Contriever. We uncover major vulnerabilities, showing retrievers favor shorter documents, early positions, repeated entities, and literal matches, all while ignoring the answer's presence! Notably, when multiple biases combine, models exhibit catastrophic performance degradation, selecting the answer-containing document in less than 10% of cases over a synthetic biased document without the answer. Furthermore, we show that these biases have direct consequences for downstream applications like RAG, where retrieval-preferred documents can mislead LLMs, resulting in a 34% performance drop than providing no documents at all. https://huggingface.co/datasets/mohsenfayyaz/ColDeR

Knowledge-Augmented Multimodal Clinical Rationale Generation for Disease Diagnosis with Small Language Models

Shuai Niu, Jing Ma, Hongzhan Lin, Liang Bai, Zhihua Wang, Richard Yi Da Xu, Yunya Song, Xian Yang

Interpretation is critical for disease diagnosis, but existing models struggle to balance predictive accuracy with human-understandable rationales. While large language models (LLMs) offer strong reasoning abilities, their clinical use is limited by high computational costs and restricted multimodal reasoning ability. Small language models (SLMs) are efficient but lack advanced reasoning for integrating multimodal medical data. In addition, both LLMs and SLMs lack domain knowledge for trustworthy reasoning. Therefore, we propose ClinRaGen, enhancing SLMs by leveraging LLM-derived reasoning ability via rationale distillation and domain knowledge injection for trustworthy multimodal rationale generation. Key innovations include a sequential rationale distillation framework that equips SLMs with LLM-comparable multimodal reasoning abilities, and a knowledge-augmented attention mechanism that jointly unifies multimodal representation from time series and textual data in the same encoding space, enabling it to be naturally interpreted by SLMs while incorporating domain knowledge for reliable rationale generation. Experiments on real-world medical datasets show that ClinRaGen achieves state-of-the-art performance in disease diagnosis and rationale generation, demonstrating the effectiveness of combining LLM-driven reasoning with knowledge augmentation for improved interpretability.

Adaptive Tool Use in Large Language Models with Meta-Cognition Trigger

Wenjun Li, Dexun Li, Kuicai Dong, Cong Zhang, Hao Zhang, Weiwen Liu, Yasheng Wang, Ruiming Tang, Yong Liu

Large language models (LLMs) have shown remarkable emergent capabilities, transforming the execution of functional tasks by leveraging external tools for complex problems that require specialized processing or up-to-date data. While existing research expands LLMs access to diverse tools (e.g., program interpreters, search engines, calculators), the necessity of using these tools is often overlooked, leading to indiscriminate tool invocation. This naive approach raises two key issues: increased latency due to unnecessary tool calls, and potential errors resulting from faulty interactions with external tools. In this paper, we introduce meta-cognition as a proxy for LLMs self-assessment of their capabilities, reflecting the model's awareness of its own limitations. Based on this, we propose MeCo, an adaptive decision-making strategy for external tool use. MeCo quantifies metacognitive scores by capturing high-level cognitive signals in the representation space, guiding when to invoke tools. Notably, MeCo is fine-tuning-free and incurs minimal cost. Experiments across multiple backbone models and benchmarks show that MeCo reliably detects LLMs' internal cognitive signals and significantly improves tool-use decision-making.

ChainEdit: Propagating Ripple Effects through Logical Rule-Guided Chain Updates

Zilu dong, Xiangqing Shen, Zinong Yang, Rui Xia

Current knowledge editing methods for large language models (LLMs) struggle to maintain logical consistency when propagating ripple effects to associated facts. We propose ChainEdit, a framework that synergizes knowledge graph-derived logical rules with LLM logical reasoning capabilities to enable systematic chain updates. By automatically extracting logical patterns from structured knowledge bases and aligning them with LLMs' internal logics, ChainEdit dynamically generates and edits logically connected knowledge clusters. Experiments demonstrate an improvement of more than 30% in logical generalization over baselines while preserving editing reliability and specificity. We further address evaluation biases in existing benchmarks through knowledge-aware protocols that disentangle external dependencies. This work establishes new state-of-the-art performance on ripple effect while ensuring internal logical consistency after knowledge editing.

The Mirage of Model Editing: Revisiting Evaluation in the Wild

Wanli Yang, Fei Sun, Jiajun Tan, Xinyu Ma, Qi Cao, Dawei Yin, Huawei Shen, Xueqi Cheng

Despite near-perfect results reported in the literature, the effectiveness of model editing in real-world applications remains unclear. To bridge this gap, we introduce QAEdit, a new benchmark aligned with widely used question answering (QA) datasets, and WILD, a task-agnostic evaluation framework designed to better reflect real-world usage of model editing. Our single editing experiments show that current editing methods perform substantially worse than previously reported (38.5% vs. 96.8%). We demonstrate that it stems from issues in the synthetic evaluation practices of prior work. Among them, the most severe is the use of teacher forcing during testing, which leaks both content and length of the ground truth, leading to overestimated performance. Furthermore, we simulate practical deployment by sequential editing, revealing that current approaches fail drastically with only 1000 edits. This work calls for a shift in model editing research toward rigorous evaluation and the development of robust, scalable methods that can reliably update knowledge in LLMs for real-world use.

MMBoundary: Advancing MLLM Knowledge Boundary Awareness through Reasoning Step Confidence Calibration

Zhitao He, Sandeep Polisetty, Zhiyuan Fan, Shujin Wu, Yuchen Huang, Yi R. Fung

In recent years, multimodal large language models (MLLMs) have made significant progress but continue to face inherent challenges in multimodal reasoning, which requires multi-level (e.g., perception, reasoning) and multi-granular (e.g., multi-step reasoning chain) advanced inferencing. Prior work on estimating model confidence tends to focus on the overall response for training and calibration, but fails to assess confidence in each reasoning step, leading to undesirable hallucination snowballing. In this work, we present MMBoundary, a novel framework that advances the knowledge boundary awareness of MLLMs through reasoning step confidence calibration. To achieve this, we propose to incorporate complementary textual and cross-modal self-rewarding signals to estimate confidence at each step of the MLLM reasoning process. In addition to supervised fine-tuning MLLM on this set of self-rewarding confidence estimation signal for initial confidence expression warm-up, we introduce a reinforcement learning stage with multiple reward functions for further aligning model knowledge and calibrating confidence at each reasoning step, enhancing reasoning chain self-correction. Empirical results show that MMBoundary significantly outperforms existing methods across diverse domain datasets and metrics, achieving an average of 7.5% reduction in multimodal confidence calibration errors and up to 8.3% improvement in task performance.

Revealing the Deceptiveness of Knowledge Editing: A Mechanistic Analysis of Superficial Editing

Jiakuan Xie, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao

Knowledge editing, which aims to update the knowledge encoded in language models, can be deceptive. Despite the fact that many existing knowledge editing algorithms achieve near-perfect performance on conventional metrics, the models edited by them are still prone to generating original knowledge. This paper introduces the concept of "**superficial editing**" to describe this phenomenon. Our comprehensive evaluation reveals that this issue presents a significant challenge to existing algorithms. Through systematic investigation, we identify and validate two key factors contributing to this issue: (1) the residual stream at the last subject position in earlier layers and (2) specific attention modules in later layers. Notably, certain attention heads in later layers, along with specific left singular vectors in their output matrices, encapsulate the original knowledge and exhibit a causal relationship with superficial editing. Furthermore, we extend our analysis to the task of superficial unlearning, where we observe consistent patterns in the behavior of specific attention heads and their corresponding left singular vectors, thereby demonstrating the robustness and broader applicability of our methodology and conclusions. Our code is available at https://github.com/jiakuan929/superficial-editing.

From Human Reading to NLM Understanding: Evaluating the Role of Eye-Tracking Data in Encoder-Based Models

Luca Dini, Lucia Domenichelli, Dominique Brunato, Felice Dell'Orletta

Cognitive signals, particularly eye-tracking data, offer valuable insights into human language processing. Leveraging eye-gaze data from the Ghent Eye-Tracking Corpus, we conducted a series of experiments to examine how integrating knowledge of human reading behavior impacts Neural Language Models (NLMs) across multiple dimensions: task performance, attention mechanisms, and the geometry of their embedding space. We explored several fine-tuning methodologies to inject eye-tracking features into the models. Our results reveal that incorporating these features does not degrade downstream task performance, enhances alignment between model attention and human attention patterns, and compresses the geometry of the embedding space.

ICR Probe: Tracking Hidden State Dynamics for Reliable Hallucination Detection in LLMs

Zhenliang Zhang, Xinyu Hu, Huixuan Zhang, Junzhe Zhang, Xiaojun Wan

Large language models (LLMs) excel at various natural language processing tasks, but their tendency to generate hallucinations undermines their reliability. Existing hallucination detection methods leveraging hidden states predominantly focus on static and isolated representations, overlooking their dynamic evolution across layers, which limits efficacy. To address this limitation, we shift the focus to the hidden state update process and introduce a novel metric, the **ICR** Score (**I**nformation **C**ontribution to **R**esidual Stream), which quantifies the contribution of modules to the hidden states' update. We empirically validate that the ICR Score is effective and reliable in distinguishing hallucinations. Building on these insights, we propose a hallucination detection method, the ICR Probe, which captures the cross-layer evolution of hidden states. Experimental results show that the ICR Probe achieves superior performance with significantly fewer parameters. Furthermore, ablation studies and case analyses offer deeper insights into the underlying mechanism of this method, improving its interpretability.

BehaviorBox: Automated Behavioral Comparison of Language Models

Lindia Tjuatja, Graham Neubig

Language model evaluation is a daunting task: prompts are brittle, corpus-level perplexities are vague, and the choice of benchmarks are endless. Finding examples that show meaningful, generalizable differences between two LMs is crucial to understanding where one model succeeds and another fails. Can this process be done automatically? In this work, we propose methodology for automated comparison of language models that uses performance-aware contextual embeddings to find fine-grained features of text where one LM outperforms another. Our method, which we name BehaviorBox, extracts coherent features that demonstrate differences with respect to the ease of generation between two LMs. Specifically, BehaviorBox finds features that describe groups of words in fine-grained contexts, such as "conditional 'were' in the phrase 'if you were'" and "exclamation marks after emotional statements", where one model outperforms another within a particular datatset. We apply BehaviorBox to compare models that vary in size, model family, and post-training, and enumerate insights into specific contexts that illustrate meaningful differences in performance which cannot be found by measures such as corpus-level perplexity alone.

Using Shapley interactions to understand how models use structure

Diganta Misra, Divyansh Singhvi, Andrej Erkelens, Raghav Jain, Isabel Papadimitriou, Naomi Saphra

Language is an intricately structured system, and a key goal of NLP interpretability is to provide methodological insights for understanding how language models internally represent this structure. In this paper, we use Shapley Taylor interaction indices (STII) in order to examine how language and speech models internally relate and structure their inputs. Pairwise Shapley interactions give us an attribution measure of how much two inputs work together to influence model outputs beyond if we linearly added their independent influences, providing a view into how models encode structural interactions between inputs. We relate the interaction patterns in models to three underlying linguistic structures: syntactic structure, non-compositional semantics, and phonetic interaction. We find that autoregressive text models encode interactions that correlate with the syntactic proximity of inputs, and that both autoregressive and masked models encode nonlinear interactions in idiomatic phrases with non-compositional semantics. Our speech results show that inputs are more entangled for pairs where a neighboring consonant is likely to influence a vowel or approximant, showing that models encode the phonetic interaction needed for extracting discrete phonemic representations.

Adversarial Tokenization

Renato Geh, Zilei Shao, Guy Van den Broeck

Current LLM pipelines account for only one possible tokenization for a given string, ignoring exponentially many alternative tokenizations during training and inference. For example, the \texttt{Llama3} standard tokenization of penguin is \texttt{[p,enguin]}, yet \texttt{[peng,uin]} is another perfectly valid alternative. In this paper, we show that despite LLMs being trained solely on one tokenization, they still retain semantic understanding of other tokenizations, raising questions about their implications in LLM safety. Put succinctly, we answer the following question: can we adversarially tokenize an obviously malicious string to evade safety and alignment restrictions? We show that not only is adversarial tokenization an effective yet previously neglected axis of attack, but it is also competitive against existing state-of-the-art adversarial approaches without changing the text of the harmful request. We empirically validate this exploit across three state-of-the-art LLMs and adversarial datasets, revealing a previously unknown vulnerability in subword models.

Neuron Empirical Gradient: Discovering and Quantifying Neurons' Global Linear Controllability

Xin Zhao, Zehui Jiang, Naoki Yoshinaga

While feed-forward neurons in pre-trained language models (PLMs) can encode knowledge, past research targeted a small subset of neurons that heavily influence outputs. This leaves the broader role of neuron activations unclear, limiting progress in areas like knowledge editing. We uncover a global linear relationship between neuron activations and outputs using neuron interventions on a knowledge probing dataset. The gradient of this linear relationship, which we call the **neuron empirical gradient (NEG)**, captures how changes in activations affect predictions. To compute NEG efficiently, we propose **NeurGrad**, enabling large-scale analysis of neuron behavior in PLMs. We also show that NEG effectively captures language skills across diverse prompts through skill neuron probing. Experiments on **MCEval8k**, a multi-genre multiple-choice knowledge benchmark, support NEG's ability to represent model knowledge. Further analysis highlights the key properties of NEG-based skill representation: efficiency, robustness, flexibility, and interdependency. Code and data are released.

Decoding Knowledge Attribution in Mixture-of-Experts: A Framework of Basic-Refinement Collaboration and Efficiency Analysis

Junzhuo Li, Bo Wang, Xiuze Zhou, Peijie Jiang, Jia Liu, Xuming Hu

The interpretability of Mixture-of-Experts (MoE) models, especially those with heterogeneous designs, remains underexplored. Existing attribution methods for dense models fail to capture dynamic routing-expert interactions in sparse MoE architectures. To address this issue, we propose a cross-level attribution algorithm to analyze sparse MoE architectures (Qwen 1.5-MoE, OLMoE, Mixtral-8x7B) against dense models (Qwen 1.5-7B, Llama-7B, Mistral-7B). Results show MoE models achieve 31\% higher per-layer efficiency via a "mid-activation, late-amplification" pattern: early layers screen experts, while late layers refine knowledge collaboratively. Ablation studies reveal a "basic-refinement" framework—shared experts handle general tasks (entity recognition), while routed experts specialize in domain-specific processing (geographic attributes). Semantic-driven routing is evidenced by strong correlations between attention heads and experts (r=0.68), enabling task-aware coordination. Notably, architectural depth dictates robustness: deep Qwen-MoE mitigates expert failures (e.g., 43% MRR drop in geographic tasks when blocking top-10 experts) through shared expert redundancy, whereas shallow Olmoe suffers severe degradation (76% drop). Task sensitivity further guides design: core-sensitive tasks (geography) require concentrated expertise, while distributed-tolerant tasks (object attributes) leverage broader participation. These insights advance MoE interpretability, offering principles to balance efficiency, specialization, and robustness.

Enhancing Chain-of-Thought Reasoning with Critical Representation Fine-tuning

Chenxi Huang, Shaotian Yan, Liang Xie, Binbin Lin, Sinan Fan, Yue Xin, Deng Cai, Chen Shen, Jieping Ye

Representation Fine-tuning (ReFT), a recently proposed Parameter-Efficient Fine-Tuning (PEFT) method, has attracted widespread attention for significantly improving parameter efficiency by editing representation space alone. In this work, we investigate applying ReFT to complex reasoning tasks. However, directly using the native ReFT method, which modifies fixed representations at the beginning and end of each layer, yields suboptimal performance, as these fixed-position representations have uncertain impact on the outputs. We observe that, in complex reasoning tasks, there often exist certain critical representations. These representations either integrate significant information from preceding layers or regulate subsequent layer representations. Through layer-by-layer propagation, they exert a substantial influence on the final output. Naturally, fine-tuning these critical representations has the potential to greatly enhance reasoning performance. Building upon these insights, we propose **C**ritical **R**epresentation **F**ine-**T**uning (CRFT), a novel method that identifies and optimizes these critical representations through information flow analysis. CRFT operates within a supervised learning framework, dynamically optimizing critical representations in a low-rank linear subspace while freezing the base model. The effectiveness and efficiency of our method are validated across eight benchmarks for arithmetic and commonsense reasoning, using LLaMA and Mistral model families. Notably, our method improves the accuracy of LLaMA-2-7B and ReFT by 18.2 and 3.8, respectively, on GSM8K, while using only 0.016 of the model parameters, significantly less than other PEFT methods. Furthermore, our method also adapts effectively to few-shot settings, boosting one-shot accuracy by 16.4. Our work highlights the untapped potential of representation-level optimization for CoT reasoning, offering a lightweight yet powerful alternative to traditional PEFT methods.

Tracing and Dissecting How LLMs Recall Factual Knowledge for Real World Questions

Yiqun Wang, Chaoqun Wan, Sile Hu, Yonggang Zhang, Xiang Tian, Yaowu Chen, Xu Shen, Jieping Ye

Recent advancements in large language models (LLMs) have shown promising ability to perform commonsense reasoning, bringing machines closer to human-like understanding. However, deciphering the internal reasoning processes of LLMs remains challenging due to the complex interdependencies among generated tokens, especially in practical question-answering. In this study, we introduce a two-dimensional analysis framework—comprising token back-tracing and individual token decoding—to uncover how LLMs conduct factual knowledge recall. Through explanatory analysis of three typical reasoning datasets, we identify a consistent three-phase pattern: Subject Augmentation and Broadcasting, Object Retrieval and Reranking, and Conclusion Fusion and Generation. Our findings reveal that LLMs do not lack relevant knowledge but struggle to select the most accurate information based on context during the retrieval and rerank phase. Leveraging these findings, we apply representation engineering and selective fine-tuning to target specific modules responsible for retrieval and rerank errors. Experimental results show large improvements in response accuracy for both in-domain and out-of-domain settings, validating the rationality of the interpreting result.

Prediction Hubs are Context-Informed Frequent Tokens in LLMs

Beatrix Miranda Ginn Nielsen, Iuri Macocco, Marco Baroni

Hubness, the tendency for a few points to be among the nearest neighbours of a disproportionate number of other points, commonly arises when applying standard distance measures to high-dimensional data, often negatively impacting distance-based analysis. As autoregressive large language models (LLMs) operate on high-dimensional representations, we ask whether they are also affected by hubness. We first prove that the only large-scale representation comparison operation performed by LLMs, namely that between context and unembedding vectors to determine continuation probabilities, is not characterized by the concentration of distances phenomenon that typically causes the appearance of nuisance hubness. We then empirically show that this comparison still leads to a high degree of hubness, but the hubs in this case do not constitute a disturbance. They are rather the result of context-modulated frequent tokens often appearing in the pool of likely candidates for next token prediction. However, when other distances are used to compare LLM representations, we do not have the same theoretical guarantees, and, indeed, we see nuisance hubs appear. There are two main takeaways. First, hubness, while omnipresent in high-dimensional spaces, is not a negative property that needs to be mitigated when LLMs are being used for next token prediction. Second, when comparing representations from LLMs using Euclidean or cosine distance, there is a high risk of nuisance hubs and practitioners should use mitigation techniques if relevant.

Representation Bending for Large Language Model Safety

Ashkan Yousefpour, Taeheon Kim, Ryan Sungmo Kwon, Seungbeen Lee, Wonje Jeung, Seungju Han, Alvin Wan, Harrison Ngan, Youngjae Yu, Jonghyun Choi

Large Language Models (LLMs) have emerged as powerful tools, but their inherent safety risks -- ranging from harmful content generation to broader societal harms -- pose significant challenges. These risks can be amplified by the recent adversarial attacks, fine-tuning vulnerabilities, and the increasing deployment of LLMs in high-stakes environments. Existing safety-enhancing techniques, such as fine-tuning with human feedback or adversarial training, are still vulnerable as they address specific threats and often fail to generalize across unseen attacks, or require manual system-level defenses. This paper introduces RepBend, a novel approach that fundamentally disrupts the representations underlying harmful behaviors in LLMs, offering a scalable solution to enhance (potentially inherent) safety. RepBend brings the idea of activation steering -- simple vector arithmetic for steering model's behavior during inference -- to loss-based fine-tuning. Through extensive evaluation, RepBend achieves state-of-the-art performance, outperforming prior methods such as Circuit Breaker, RMU, and NPO, with up to 95% reduction in attack success rates across diverse jailbreak benchmarks, all with negligible reduction in model usability and general capabilities.

Memorization Inheritance in Sequence-Level Knowledge Distillation for Neural Machine Translation

Verna Dankers, Vikas Raunak

In this work, we explore how instance-level memorization in the teacher Neural Machine Translation (NMT) model gets inherited by the student model in sequence-level knowledge distillation (SeqKD). We find that despite not directly seeing the original training data, students memorize more than baseline models (models of the same size, trained on the original data)—3.4% for exact matches and 57% for extractive memorization—and show increased hallucination rates. Further, under this SeqKD setting, we also characterize how students behave on specific training data subgroups, such as subgroups with low quality or specific counterfactual memorization (CM) scores, and find that students exhibit greater denoising on low-quality subgroups. Finally, we propose a modification to SeqKD named Adaptive-SeqKD, which intervenes in SeqKD to reduce memorization and hallucinations. Overall, we recommend caution when applying SeqKD: students inherit both their teachers' superior performance and their fault modes, thereby requiring active monitoring.

Library-Like Behavior In Language Models is Enhanced by Self-Referencing Causal Cycles

Munachiso S Nwadike, Zangir Iklassov, Toluwani Aremu, Tatsuya Hiraoka, Benjamin Heinzerling, Velibor Bojkovic, Hilal AlQuabeh, Martin Takáč, Kentaro Inui

We introduce the concept of the \textit{self-referencing causal cycle} (abbreviated \textit{ReCall})—a mechanism that enables large language models (LLMs) to bypass the limitations of unidirectional causality, which underlies a phenomenon known as the \textit{reversal curse}. When an LLM is prompted with sequential data, it often fails to recall preceding context. For example, when we ask an LLM to recall the line preceding "O say does that star-spangled banner yet wave" in the U.S. National Anthem, it often fails to correctly return "Gave proof through the night that our flag was still there"—this is due to the reversal curse. It occurs because language models such as ChatGPT and Llama generate text based on preceding tokens, requiring facts to be learned and reproduced in a consistent token order. While the reversal curse is often viewed as a limitation, we offer evidence of an alternative view: it is not always an obstacle in practice. We find that \textit{ReCall} is driven by what we designate as \textit{cycle tokens}—sequences that connect different parts of the training data, enabling recall of preceding tokens from succeeding ones. Through rigorous probabilistic formalization and controlled experiments, we demonstrate how the cycles they induce influence a model's ability to reproduce information. To facilitate reproducibility, we provide our code and experimental details at https://anonymous.4open.science/r/remember-B0B8/.

Shaping the Safety Boundaries: Understanding and Defending Against Jailbreaks in Large Language Models

Lang Gao, Jiahui Geng, Xiangliang Zhang, Preslav Nakov, Xiuying Chen

Jailbreaking in Large Language Models (LLMs) is a major security concern as it can deceive LLMs into generating harmful text. However, understanding of how jailbreaking works remains limited, hindering the development of effective defense strategies. To address this issue, we conduct a large-scale analysis of seven different jailbreak methods and identify that disagreements among methods stem from insufficient observation samples. We introduce the concept of a safety boundary and discover that jailbreaks shift harmful activations outside this boundary, where LLMs become less sensitive to harmful information. Our analysis reveals that low and middle layers play a critical role in these shifts, while deeper layers have a lesser impact. Building on these insights, we propose a novel defense mechanism called Activation Boundary Defense (ABD), which adaptively constrains activations within the safety boundary. To enhance its effectiveness, we use Bayesian optimization to selectively apply the defense to the low and middle layers. Experiments on several benchmark datasets demonstrate that ABD achieves an average Defense Success Rate (DSR) of over 98% against various jailbreak attacks, with less than a 2% impact on the model’s general capabilities.

Exploring the Impact of Instruction-Tuning on LLM's Susceptibility to Misinformation

Kyubeen Han, Junseo Jang, Hongjin Kim, Geunyeong Jeong, Harksoo Kim

Instruction-tuning enhances the ability of large language models (LLMs) to follow user instructions more accurately, improving usability while reducing harmful outputs. However, this process may increase the model's dependence on user input, potentially leading to the unfiltered acceptance of misinformation and the generation of hallucinations. Existing studies primarily highlight that LLMs are receptive to external information that contradict their parametric knowledge, but little research has been conducted on the direct impact of instruction-tuning on this phenomenon. In our study, we investigate the impact of instruction-tuning on LLM susceptibility to misinformation. Our analysis reveals that instruction-tuned LLMs are significantly more likely to accept misinformation when it is presented by the user. A comparison with base models shows that instruction-tuning increases reliance on user-provided information, shifting susceptibility from the assistant role to the user role. Furthermore, we explore additional factors influencing misinformation susceptibility, such as the role of the user in prompt structure, misinformation length, and the presence of warnings in the system prompt. Our findings underscore the need for systematic approaches to mitigate unintended consequences of instruction-tuning and enhance the reliability of LLMs in real-world applications.

The Causal Effect of Merge Operations in Bottom-up Tokenisers

Pietro Lesci, Clara Meister, Thomas Hofmann, Andreas Vlachos, Tiago Pimentel

Modern language models are typically trained over subword sequences, but ultimately define probabilities over character-strings. Ideally, the choice of the tokeniser---which maps character-strings to subwords---should not affect the probability assigned to the underlying character-string; in practice, it does. We define this mismatch as **tokenisation bias**. In this work, we quantify one particular type of tokenisation bias: the effect of including or not a subword (e.g., \langle hello \rangle) in a tokeniser's vocabulary on the probability a trained model assigns to the corresponding characters (i.e., "hello"). Estimating this effect is challenging because each model is trained with only one tokeniser. We address this by framing tokenisation bias as a causal effect and estimating it using the regression discontinuity design. Specifically, we exploit the fact that tokenisation algorithms rank subwords and add the first K to a tokeniser's vocabulary, where K is an arbitrary cutoff point. As such, we can estimate a causal effect by comparing similar subwords around this cutoff. Experimentally, we find that tokenisation consistently affects models' outputs across scales, vocabularies, and tokenisers. Notably, a subword's presence in a small model's vocabulary may increase its characters' probability by up to 17 times, highlighting tokenisation as a key design choice in language modelling.

Language Model Probabilities are Not Calibrated in Numeric Contexts

Charles Lovering, Michael Krumdick, Viet Dac Lai, Varshini Reddy, Seth Ebner, Nilesh Kumar, Rik Koncel-Kedziorski, Chris Tanner

Some statements have one well-defined continuation (e.g., "the Eiffel Tower is in [Paris]"), whereas others have a natural distribution over multiple options (e.g., "the weighted coin flip was [Heads/Tails]."). We argue that language model (LM) outputs should capture these natural distributions. Our work specifically tests whether LM output probabilities are calibrated to numeric information within their textual contexts. For example, if the context (the prompt) concerns two equally likely options (e.g., heads or tails for a fair coin), the LM output probabilities should also be equal. Likewise, in a context with nonuniformly likely events (e.g., rolling a pair with two dice) an LM should output proportionate probabilities. However, we find that even in simple settings, the best LMs (1) are poorly calibrated and (2) have systematic biases: artifacts like word identity, word order, and word frequency all impact calibration. For example, `gpt-4o-mini` often picks the first of two options presented in the prompt regardless of the options' implied likelihoods, whereas `Llama-3.1-8B` picks the second. Models do not allocate probability mass among valid options in a calibrated manner.

Reconsidering LLM Uncertainty Estimation Methods in the Wild

Duygu Nur Yaldiz, Yavuz Faruk Bakman, Sungmin Kang, Tuo Zhang, Baturalp Buyukates, Sai Praneeth Karimireddy, Salman Avestimehr

Large Language Model (LLM) Uncertainty Estimation (UE) methods have become a crucial tool for detecting hallucinations in recent years. While numerous UE methods have been proposed, most existing studies evaluate them in isolated short-form QA settings using threshold-independent metrics such as AUROC or PRR. However, real-world deployment of UE methods introduces several challenges. In this work, we systematically examine four key aspects of deploying UE methods in practical settings. Specifically, we assess (1) the sensitivity of UE methods to decision threshold selection, (2) their robustness to query transformations such as typos, adversarial prompts, and prior chat history, (3) their applicability to long-form generation, and (4) strategies for handling multiple UE scores for a single query. Our evaluations on 19 UE methods reveal that most of them are highly sensitive to threshold selection when there is a distribution shift in the calibration dataset. While these methods generally exhibit robustness against previous chat history and typos, they are significantly vulnerable to adversarial prompts. Additionally, while existing UE methods can be adapted for long-form generation through various strategies, there remains considerable room for improvement. Lastly, ensembling multiple UE scores at test time provides a notable performance boost, which highlights its potential as a practical improvement strategy. Code is available at: https://github.com/duygunuryldz/uncertainty_in_the_wild.

PlanningArena: A Modular Benchmark for Multidimensional Evaluation of Planning and Tool Learning

Zihan Zheng, Tianle Cui, Chuwen Xie, Jiahui Pan, Qianglong Chen, Lewei He

One of the research focuses of large language models (LLMs) is the ability to generate action plans. Recent studies have revealed that the performance of LLMs can be significantly improved by integrating external tools. Based on this, we propose a benchmark framework called PlanningArena, which aims to simulate real application scenarios and provide a series of apps and API tools that may be involved in the actual planning process. This framework adopts a modular task structure and combines user portrait analysis to evaluate the ability of LLMs in correctly selecting tools, logical reasoning in complex scenarios, and parsing user information. In addition, we deeply diagnose the task execution effect of LLMs from both macro and micro levels. The experimental results show that even the most outstanding GPT-4o and DeepSeekV3 models only achieved a total score of 56.5% and 41.9% in PlanningArena, respectively, indicating that current LLMs still face challenges in logical reasoning, context memory, and tool calling when dealing with different structures, scenarios, and their complexity. Through this benchmark, we further explore the path to optimize LLMs to perform planning tasks.

Session 12: Language Modeling (Wednesday 30 July, 11:00-12:30, Hall 4/5)

FRAME: Boosting LLMs with A Four-Quadrant Multi-Stage Pretraining Strategy

Xuemiao Zhang, Feiyu Duan, Xu Liangyu, Yongwei Zhou, Sirui Wang, Rongxiang Weng, Jingang Wang, Xunliang Cai

Large language models (LLMs) have significantly advanced human language understanding and generation, with pretraining data quality and organization being crucial to their performance. Multi-stage pretraining is a promising approach, but existing methods often lack quantitative criteria for data partitioning and instead rely on intuitive heuristics. In this paper, we propose the novel Four-quadRAnt Multi-stage prEtraining strategy (FRAME), guided by the established principle of organizing the pretraining process into four stages to achieve significant loss reductions four times. This principle is grounded in two key findings: first, training on high Perplexity (PPL) data followed by low PPL data, and second, training on low PPL difference (PD) data followed by high PD data, both causing the loss to drop significantly twice and performance enhancements. By partitioning data into four quadrants and strategically organizing them, FRAME achieves a remarkable 16.8% average improvement over random across MMLU and CMMLU for the 3B model, effectively boosting LLM performance.

SelectLLM: Query-Aware Efficient Selection Algorithm for Large Language Models

Kaushal Kumar Maurya, KV Aditya Srivatsa, Ekaterina Kochmar

Large language models (LLMs) have been widely adopted due to their remarkable performance across various applications, driving the accelerated development of a large number of diverse models. However, these individual LLMs show limitations in generalization and performance on complex tasks due to inherent training biases, model size constraints, and the quality or diversity of pre-training datasets. A promising direction is to efficiently harness the diverse capabilities of LLMs to overcome these individual limitations. To address these limitations, we introduce a novel LLM selection algorithm called SelectLLM, which efficiently directs input queries to the most suitable subset of LLMs from a large pool, ensuring that the selected models collectively provide accurate responses. SelectLLM employs a multi-label classifier and policy based on the classifier's predictions and confidence scores in selecting an optimal, query-aware, and lightweight subset of LLMs. Our findings indicate that the proposed model outperforms existing ensemble-based baselines and achieves competitive performance with similarly sized top-performing LLMs while maintaining efficiency. Specifically, it achieves a huge reduction in inference latency on two challenging reasoning benchmarks: 13% on GSM8K and 70% on MMLU, compared to the top-performing baseline. Also, we establish a theoretical upper bound by an Oracle with LLMs and perform an in-depth linguistic analysis to understand the performance gap between the Oracle and SelectLLM.

Disentangling Reasoning Tokens and Boilerplate Tokens For Language Model Fine-tuning

Ziang Ye, Zhenru Zhang, Yang Zhang, jianxing ma, Junyang Lin, Fuli Feng

When using agent-task datasets to enhance agent capabilities for Large Language Models (LLMs), current methodologies often treat all tokens within a sample equally. However, we argue that tokens serving different roles—specifically, reasoning tokens versus boilerplate tokens (e.g., those governing output format)—differ significantly in importance and learning complexity, necessitating their disentanglement and distinct treatment. To address this, we propose a novel Shuffle-Aware Discriminator (SHAD) for adaptive token discrimination. SHAD classifies tokens by exploiting predictability differences observed after shuffling input-output combinations across samples: boilerplate tokens, due to their repetitive nature among samples, maintain predictability, whereas reasoning tokens do not. Using SHAD, we propose the Reasoning-highlighted Fine-Tuning (RFT) method, which adaptively emphasizes reasoning tokens during fine-tuning, yielding notable performance gains over common Supervised Fine-Tuning (SFT).

Preference Curriculum: LLMs Should Always Be Pretrained on Their Preferred Data

Xuemiao Zhang, Xu Liangyu, Feiyu Duan, Yongwei Zhou, Sirui Wang, Rongxiang Weng, Jingang Wang, Xunliang Cai

Large language models (LLMs) generally utilize a consistent data distribution throughout the pretraining process. However, as the model's capability improves, it is intuitive that its data preferences dynamically change, indicating the need for pretraining with different data at various training stages. To achieve it, we propose the Perplexity Difference (PD) based Preference Curriculum learning (PDPC) framework, which always perceives and uses the data preferred by LLMs to train and boost them. First, we introduce the PD metric to quantify the difference in how challenging a sample is for weak versus strong models. Samples with high PD are more challenging for weak models to learn and are more suitable to be arranged in the later stage of pretraining. Second, we propose the preference function to approximate and predict the data preference of the LLM at any training step, so as to complete the arrangement of the dataset offline and ensure continuous training without interruption. Experimental results on 1.3B and 3B models demonstrate that PDPC significantly surpasses baselines. Notably, the 3B model trained on 1T tokens achieves an increased average accuracy of over 8.1% across MMLU and CMMLU.

CARE-STaR: Constraint-aware Self-taught Reasoner

Zhiliang Li, Bo Tang, Yijun Niu, Beihong Jin, Qiwen Shi, Yuchen Feng, Zhiyu li, Jie Hu, mingchuan yang, Feiyu Xiong

In real-world applications, large language models (LLMs) often need to handle diverse and complex instructions. Specifically, when instructions are subject to multiple constraints, some of which are somewhat ambiguous, LLMs often fail to produce answers that satisfy all constraints, limiting their effectiveness in various tasks. To address this challenge, we examine the different constraints in the instructions and discover that the difficulty of determining whether an answer meets a constraint varies widely, from extremely straightforward to exceptionally perplexing. Correspondingly, we propose to assign constraints to different constraint levels. Furthermore, inspired by chain-of-thought (CoT) and self-taught reasoner (STaR), we propose a two-stage method named CARE-STaR (Constraint-AwaRE STaR). Our method distinguishes constraints within instructions by generating different CoTs and guides LLMs to autonomously learn optimal answers by setting the positive rewards to the CoTs that are beneficial to generating accurate answers and iteratively optimizing these answers. We have conducted extensive experiments on three instruction-following benchmarks, taking three existing LLMs as base LLMs, respectively. Experimental results indicate that our method substantially enhances the capability of these LLMs to handle complex instructions, outperforming supervised fine-tuning (SFT). Our code is available at https://github.com/lzl0124/carestar.

A Joint Optimization Framework for Enhancing Efficiency of Tool Utilization in LLM Agents

Bin Wu, Edgar Meij, Emine Yilmaz

Large Language Models (LLMs) augmented with external tools have demonstrated remarkable capabilities in complex problem solving. Existing efforts for tool utilization typically involve an LLM agent that contains instructions on using the description of the available tools to determine and call the tools required to solve the problem. Inference Scaling techniques, such as chain-of-thought and tree-of-thought reasoning, are commonly used but require significant computational overhead and rendering such methods impractical in real-world applications. In this work, we recognize and formalize the critical role of instructions provided in agent prompts and tool descriptions---collectively referred to as context---and show that incomplete context is one of the reasons for this computational overhead. To fill this efficiency gap, we propose an optimization framework that jointly refines both the instructions provided in the agent prompt and tool description, enhancing their interaction. Experiments on StableToolBench and RestBench demonstrate that our optimized agents achieve superior efficiency while maintaining effectiveness. Our findings underscore the critical role of context optimization in improving LLM agents for tool utilization, paving the way for more responsive and cost-effective LLM agents.

Unilogit: Robust Machine Unlearning for LLMs Using Uniform-Target Self-Distillation

Stefan Vasilev, Christian Herold, Baohao Liao, Seyyed Hadi Hashemi, Shahram Khadivi, Christof Monz

This paper introduces Unilogit, a novel self-distillation method for machine unlearning in Large Language Models. Unilogit addresses the challenge of selectively forgetting specific information while maintaining overall model utility, a critical task in compliance with data privacy regulations like GDPR. Unlike prior methods that rely on static hyperparameters or starting model outputs, Unilogit dynamically adjusts target logits to achieve a uniform probability for the target token, leveraging the current model's outputs for more accurate self-distillation targets. This approach not only eliminates the need for additional hyperparameters but also enhances the model's ability to approximate the golden targets. Extensive experiments on public benchmarks and an in-house e-commerce dataset demonstrate Unilogit's superior performance in balancing forget and retain objectives, outperforming state-of-the-art methods such as NPO and UnDIAL. Our analysis further reveals Unilogit's robustness across various scenarios, highlighting its practical applicability and effectiveness in achieving efficacious machine unlearning.

Towards Safety Reasoning in LLMs: AI-agentic Deliberation for Policy-embedded CoT Data Creation

Tharindu Kumarage, Ninareh Mehrabi, Anil Ramakrishna, Xinyan Zhao, Richard Zemel, Kai-Wei Chang, Aram Galstyan, Rahul Gupta, Charith Peris

Safety reasoning is a recent paradigm where LLMs reason over safety policies before generating responses, thereby mitigating limitations in existing safety measures such as over-refusal and jailbreak vulnerabilities. However, implementing this paradigm is challenging due to the resource-intensive process of creating high-quality policy-embedded chain-of-thought (CoT) datasets while ensuring reasoning remains accurate and free from hallucinations or policy conflicts. To tackle this, we propose AIDSAFE: Agentic Iterative Deliberation for Safety Reasoning, a novel data generation recipe that leverages multi-agent deliberation to iteratively expand reasoning on safety policies. A data refiner stage in AIDSAFE ensures high-quality outputs by eliminating repetitive, redundant, and deceptive thoughts. AIDSAFE-generated CoTs provide a strong foundation for supervised fine-tuning (SFT)-based safety training. Additionally, to address the need of preference data in alignment stages, such as DPO training, we introduce a supplemental recipe that uses belief augmentation to create distinct selected and rejected CoT samples. Our evaluations demonstrate that AIDSAFE-generated CoTs achieve superior policy adherence and reasoning quality. Consequently, we show that fine-tuning open-source LLMs on these CoTs can significantly improve safety generalization and jailbreak robustness while maintaining acceptable utility and over-refusal accuracy.

HG-InsightLog: Context Prioritization and Reduction for Question Answering with Non-Natural Language Construct Log Data

Supriya Bajpai, Athira Gopal, Chandrakant Harjpal, Niraj Kumar

Modern IT systems generate vast amounts of log data, which pose challenges for Large Language Models (LLMs) due to their large size, irrelevant entries, and non-Natural Language (non-NL) construct (e.g., domain-specific jargon, error codes, file paths, and abbreviations). Traditional methods like Retrieval-Augmented Generation (RAG) and GraphRAG fail to preserve temporal sequences, handle non-NL for context and entities extraction, and dynamically prioritize query-relevant context. To address these limitations, we propose HG-InsightLog, a novel framework that constructs a multi-entity temporal hypergraph representing log attribute-value pair as nodes and connecting them with hyperedges, capturing critical connections in the data. HG-InsightLog introduces a multi-step query personalization mechanism enhancing the Personalized PageRank algorithm to rank hyperedges based on query relevance and contextual centrality to priortize critical connections. Top ranked hyperedges are extracted and converted back into log formats preserving temporal order and reducing context. Experimental results across multiple datasets demonstrate its superiority over existing methods, enhancing factual, causal, and analytical reasoning. Our approach enables smaller LLMs like LLaMA-8B to perform effective log-based QA. Being model-agnostic and training-free, it scales with evolving open-source LLMs without relying on proprietary systems.

Source Attribution for Large Language Model-Generated Data

Xinyang Lu, Jingtan Wang, Zitong Zhao, Zhongxiang Dai, Chuan-Sheng Foo, See-Kiong Ng, Bryan Kian Hsiang Low

The impressive performances of Large Language Models (LLMs) and their immense potential for commercialization have given rise to serious concerns over the Intellectual Property (IP) of their training data. In particular, the synthetic texts generated by LLMs may infringe the IP of the data being used to train the LLMs. To this end, it is imperative to be able to perform source attribution by identifying the data provider who contributed to the generation of a synthetic text by an LLM. In this paper, we show that this problem can be tackled by watermarking, i.e., by enabling an LLM to generate synthetic texts with embedded watermarks that contain information about their source(s). We identify the key properties of such watermarking frameworks (e.g., source attribution accuracy, robustness against adversaries), and propose a source attribution framework that satisfies these key properties due to our algorithmic designs. Our framework enables an LLM to learn an accurate mapping from the generated texts to data providers, which sets the foundation for effective source attribution. Extensive empirical evaluations show that our framework achieves effective source attribution.

Don't Say No: Jailbreaking LLM by Suppressing Refusal

Yukai Zhou, Jian Lou, Zhijie Huang, Zhan Qin, Sibei Yang, Wenjie Wang

Ensuring the safety alignment of Large Language Models (LLMs) is critical for generating responses consistent with human values. However, LLMs remain vulnerable to jailbreaking attacks, where carefully crafted prompts manipulate them into producing toxic content. One category of such attacks reformulates the task as an optimization problem, aiming to elicit affirmative responses from the LLM. However, these methods heavily rely on predefined objectionable behaviors, limiting their effectiveness and adaptability to diverse harmful queries. In this study, we first identify why the vanilla target loss is suboptimal and then propose enhancements to the loss objective. We introduce \textit{DSN} (Don't Say No) attack, which combines cosine decay schedule method with refusal suppression to achieve higher success rates. Extensive experiments demonstrate that \textit{DSN} outperforms baseline attacks and achieves state-of-the-art attack success rates (ASR). \textit{DSN} also shows strong universality and transferability to unseen datasets and black-box models.

Revisiting In-Context Learning with Long Context Language Models

Jinheon Baek, Sun Jae Lee, Prakhar Gupta, Geunseob Oh, Siddharth Dalmia, Prateek Kolhar

In-Context Learning (ICL) is a technique by which language models make predictions based on examples provided in their input context. Previously, their context window size imposed a limit on the number of examples that can be shown, making example selection techniques crucial for identifying the maximally effective set of examples. However, the recent advent of Long Context Language Models (LCLMs) has significantly increased the number of examples that can be included in context, raising an important question of whether ICL performance in a many-shot regime is still sensitive to the method of sample selection. To answer this, we revisit these approaches in the context of LCLMs through extensive experiments on 18 datasets spanning 4 tasks. Surprisingly, we observe that sophisticated example selection techniques do not yield significant improvements over a simple random sample selection method. Instead, we discover that the advent of LCLMs has fundamentally shifted the challenge of ICL from that of selecting the most effective examples to that of collecting sufficient examples to fill the context window. Specifically, in certain datasets, including all available examples does not fully utilize the context window; however, by augmenting the examples in context with a simple data augmentation approach, we substantially improve ICL performance by 5%.

GAPO: Learning Preferential Prompt through Generative Adversarial Policy Optimization

Zhouhong Gu, Xingzhou Chen, Xiaoran Shi, Tao Wang, Suhang Zheng, Tianyu Li, Hongwei Feng, Yanghua Xiao

Recent advances in large language models have highlighted the critical need for precise control over model outputs through predefined constraints. While existing methods attempt to achieve this through either direct instruction-response synthesis or preferential response optimization, they often struggle with constraint understanding and adaptation. This limitation becomes particularly evident when handling fine-grained constraints, leading to either hallucination or brittle performance. We introduce Generative Adversarial Policy Optimization (GAPO), a novel framework that combines GAN-based training dynamics with an encoder-only reward model to progressively learn and adapt to increasingly complex constraints. GAPO leverages adversarial training to automatically generate training samples of varying difficulty while utilizing the encoder-only architecture to better capture prompt-response relationships. Extensive experiments demonstrate GAPO's superior performance across multiple benchmarks, particularly in scenarios requiring fine-grained constraint handling, where it significantly outperforms existing methods like PPO, DPO, and KTO. Our results suggest that GAPO's unique approach to preferential prompt learning offers a more robust and effective solution for controlling LLM outputs.

Untie the Knots: An Efficient Data Augmentation Strategy for Long-Context Pre-Training in Language Models

Junfeng Tian, Da Zheng, Yang Chen, Rui Wang, colin zhang, Debing Zhang

Large language models (LLM) have prioritized expanding the context window from which models can incorporate more information. However, training models to handle long contexts presents significant challenges. These include the scarcity of high-quality natural long-context data, the potential for performance degradation on short-context tasks, and the reduced training efficiency associated with attention mechanisms. In this paper, we introduce Untie the Knots (UtK), a novel data augmentation strategy employed during the continue pre-training phase, designed to efficiently enable LLMs to gain long-context capabilities without the need to modify the existing data mixture. In particular, we chunk the documents, shuffle the chunks, and create a complex and knotted structure of long texts; LLMs are then trained to untie these knots and identify relevant segments within seemingly chaotic token sequences. This approach greatly improves the model's performance by accurately attending to relevant information in long context and the training efficiency is also largely increased. We conduct extensive experiments on models with 7B and 72B parameters, trained on 20 billion tokens, demonstrating that UtK achieves 75% and 84.5% accurracy on RULER at 128K context length, significantly outperforming other long context strategies. The trained models will open-source for further research.

Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, Iacopo Poli

Encoder-only transformer models such as BERT offer a great performance-size tradeoff for retrieval and classification tasks with respect to larger decoder-only models. Despite being the workhorse of numerous production pipelines, there have been limited Pareto improvements to BERT since its release. In this paper, we introduce ModernBERT, bringing modern model optimizations to encoder-only models and representing a major Pareto improvement over older encoders. Trained on 2 trillion tokens with a native 8192 sequence length, ModernBERT models exhibit state-of-the-art results on a large pool of evaluations encompassing diverse classification tasks and both single and multi-vector retrieval on different domains (including code). In addition to strong downstream performance, ModernBERT is also the most speed and memory efficient encoder and is designed for inference on common GPUs.

Mixture of insighTful Experts (MoTE): The Synergy of Reasoning Chains and Expert Mixtures in Self-Alignment

Zhili Liu, Yunhao GOU, Kai Chen, Lanqing HONG, Jiahui Gao, Fei Mi, Yu Zhang, Zhenguo Li, Xin Jiang, Qun Liu, James Kwok

As the capabilities of large language models (LLMs) continue to expand, aligning these models with human values remains a significant challenge. Recent studies show that reasoning abilities contribute significantly to model safety, while integrating Mixture-of-Experts (MoE) architectures can further enhance alignment. In this work, we address a fundamental question: \textit{How to effectively incorporate reasoning abilities and MoE architectures into self-alignment process in LLMs?} We propose Mixture of insighTful Experts (MoTE), a novel framework that synergistically combines reasoning chains and expert mixtures to improve self-alignments. From a data perspective, MoTE employs a structured reasoning chain comprising four key stages: Question Analysis, Answer Guidance, Safe Answer, and Safety Checking. This approach enhances safety through multi-step reasoning and proves effective even for smaller and less powerful LLMs (e.g., 7B models). From an architectural perspective, MoTE adopts a multi-LoRA framework with step-level routing, where each expert is dedicated to a specific reasoning step. This design eliminates the need for balance losses, ensures stable training, and supports adaptive inference lengths. Experimental results demonstrate that MoTE significantly improves model safety, jailbreak resistance, and over-refusal capabilities, achieving performance comparable to OpenAI's state-of-the-art o1 model.

Towards Reward Fairness in RLHF: From a Resource Allocation Perspective

Sheng Ouyang, Yulan Hu, Ge Chen, Qingyang Li, Fuzheng Zhang, Yong Liu

Rewards serve as proxies for human preferences and play a crucial role in Reinforcement Learning from Human Feedback (RLHF). However, if these rewards are inherently imperfect, exhibiting various biases, they can adversely affect the alignment of large language models (LLMs). In this paper, we collectively define the various biases present in rewards as the problem of reward unfairness. We propose a bias-agnostic method to address the issue of reward fairness from a resource allocation perspective, without specifically designing for each type of bias, yet effectively mitigating them. Specifically, we model preference learning as a resource allocation problem, treating rewards as resources to be allocated while considering the trade-off between utility and fairness in their distribution. We propose two methods, Fairness Regularization and Fairness Coefficient, to achieve fairness in rewards. We apply our methods in both verification and reinforcement learning scenarios to obtain a fairness reward model and a policy model, respectively. Experiments conducted in these scenarios demonstrate that our approach aligns LLMs with human preferences in a more fair manner. Our data and code are available at .

Improve Safety Training of Large Language Models with Safety-Critical Singular Vectors Localization

Peijian Gu, Quan Wang, Zhendong Mao

The rapid advancement of large language models (LLMs) has brought about increased concerns regarding their safety, especially as adversaries develop jailbreak techniques to bypass LLMs' safety mechanism. Although recent work on safety training with modules such as low-rank adaptation (LoRA) to resist jailbreaks shows promise, these approaches can inadvertently degrade a model’s general utility. In this paper, we propose a novel plug-and-play method that mitigates the impact of safety training on model utility by explicitly locating and leveraging safety-critical singular vectors, which only contribute to safety, within the model’s parameter space. We quantify the safety-criticality of each singular vector as the difference of their importance for safety and utility measured by a corresponding low-rank projection. The top scored singular vectors are located as safety-critical and are used to initialize the LoRA modules within existing safety training methods in a plug-and-play manner, thereby constraining the training updates within safety-critical parameters. Additionally, we propose a dynamic rank number determination strategy to further reduce parameter overhead. Experiments on HarmBench with multiple jailbreak methods validate the effectiveness of our approach in safety training, while evaluations on several utility benchmarks demonstrate that our method successfully mitigates the adverse impact of safety training on model utility, enhancing the utility performance of the evaluated safety training baselines.

MergePrint: Merge-Resistant Fingerprints for Robust Black-box Ownership Verification of Large Language Models

Shojiro Yamabe, Futa Kai Waseda, Tsubasa Takahashi, Koki Wataoka

Protecting the intellectual property of Large Language Models (LLMs) has become increasingly critical due to the high cost of training. Model merging, which integrates multiple expert models into a single multi-task model, introduces a novel risk of unauthorized use of LLMs due to its efficient merging process. While fingerprinting techniques have been proposed for verifying model ownership, their resistance to model merging remains unexplored. To address this gap, we propose a novel fingerprinting method, MergePrint, which embeds robust fingerprints capable of surviving model merging. MergePrint enables black-box ownership verification, where owners only need to check if a model produces target outputs for specific fingerprint inputs, without accessing model weights or intermediate outputs. By optimizing against a pseudo-merged model that simulates merged behavior, MergePrint ensures fingerprints that remain detectable after merging. Additionally, to minimize performance degradation, we pre-optimize the fingerprint inputs. MergePrint pioneers a practical solution for black-box ownership verification, protecting LLMs from misappropriation via merging, while also excelling in resistance to broader model theft threats.

Dynamic Scaling of Unit Tests for Code Reward Modeling

Zeyao Ma, Xiaokang Zhang, Jing Zhang, Jifan Yu, Sijia Luo, Jie Tang

Current large language models (LLMs) often struggle to produce accurate responses on the first attempt for complex reasoning tasks like code generation. Prior research tackles this challenge by generating multiple candidate solutions and validating them with LLM-generated unit tests. The execution results of unit tests serve as reward signals to identify correct solutions. As LLMs always confidently make mistakes, these unit tests are not reliable, thereby diminishing the quality of reward signals. Motivated by the observation that scaling the number of solutions improves LLM performance, we explore the impact of scaling unit tests to enhance reward signal quality. Our pioneer experiment reveals a positive correlation between the number of unit tests and reward signal quality, with greater benefits observed in more challenging problems. Based on these insights, we propose CodeRM-8B, a lightweight yet effective unit test generator that enables efficient and high-quality unit test scaling. Additionally, we implement a dynamic scaling mechanism that adapts the number of unit tests based on problem difficulty, further improving efficiency. Experimental results show that our approach significantly improves performance across various models on three benchmarks (e.g., with gains of 18.43 for Llama3-8B and 3.42 for GPT-4o-mini on HumanEval Plus). The parameters of CodeRM-8B and corresponding training data will be available upon publication.

LLMs + Persona-Plug = Personalized LLMs

Jiongnan Liu, Yutao Zhu, Shuting Wang, Xiaochi Wei, Erxue Min, Yu Lu, Shuaiqiang Wang, Dawei Yin, Zhicheng Dou

Personalization plays a critical role in numerous language tasks and applications, since users with the same requirements may prefer diverse outputs based on their interests. This has led to the development of various personalized approaches aimed at adapting large language models (LLMs) to generate customized outputs aligned with user preferences. Some of them involve fine-tuning a unique personalized LLM for each user, which is too expensive for widespread application. Alternative approaches introduce personalization information in a plug-and-play manner by retrieving the user's relevant historical texts as demonstrations. However, this retrieval-based strategy may break the continuity of the user history and fail to capture the user's overall styles and patterns, hence leading to sub-optimal performance. To address these challenges, we propose a novel personalized LLM model, PPlug. It constructs a user-specific embedding for each individual by modeling all her historical contexts through a lightweight plug-in user embedder module. By attaching this embedding to the task input, LLMs can better understand and capture user habits and preferences, thereby producing more personalized outputs without tuning their parameters. Extensive experiments on various tasks in the language model personalization (LaMP) benchmark demonstrate that the proposed model significantly outperforms existing personalized LLM approaches.

InvestAlign: Overcoming Data Scarcity in Aligning Large Language Models with Investor Decision-Making Processes Under Herd Behavior

Huisheng Wang, Zhuoshi Pan, Hangjing Zhang, Mingxiao Liu, Hanqing Gao, H. Vicky Zhao

Aligning Large Language Models (LLMs) with investor decision-making processes under herd behavior is a critical challenge in behavioral finance, which grapples with a fundamental limitation: the scarcity of real-user data needed for Supervised Fine-Tuning (SFT). While SFT can bridge the gap between LLM outputs and human behavioral patterns, its reliance on massive authentic data imposes substantial collection costs and privacy risks. We propose **InvestAlign**, a novel framework that constructs high-quality SFT datasets by leveraging theoretical solutions to similar and simple optimal investment problems rather than the complex scenarios. Our theoretical analysis demonstrates that training LLMs with **InvestAlign**-generated data achieves faster parameter convergence than using real-user data, suggesting superior learning efficiency. Furthermore, we develop **InvestAgent**, an LLM agent fine-tuned with **InvestAlign**, which shows significantly closer alignment to real-user data than pre-SFT models in both simple and complex investment problems. This highlights our proposed **InvestAlign** as a promising approach with the potential to address complex optimal investment problems and align LLMs with investor decision-making processes under herd behavior. Our code is publicly available at https://github.com/thu-social-network-research-group/InvestAlign.

DAPE V2: Process Attention Score as Feature Map for Length Extrapolation

Chuanyang Zheng, Yihang Gao, Han Shi, Jing Xiong, Jiankai Sun, Jingyao Li, Minbin Huang, Xiaozhe Ren, Michael Ng, Xin Jiang, Zhenguo Li, Yu Li

The attention mechanism is a fundamental component of the Transformer model, contributing to interactions among distinct tokens. In general, the attention scores are determined simply by the key-query products. However, this work's occasional trial (combining DAPE and NoPE) of including additional MLPs on attention scores without position encoding indicates that the classical key-query multiplication may limit the performance of Transformers. In this work, we conceptualize attention as a feature map and apply the convolution operator (for neighboring attention scores across different heads) to mimic the processing methods in computer vision. Specifically, **the main contribution of this paper is identifying and interpreting the Transformer length extrapolation problem as a result of the limited expressiveness of the naive query and key dot product, and we successfully translate the length extrapolation issue into a well-understood feature map processing problem**, which is called Convolutional Data-Adaptive Position Encoding (CDAPE). The novel insight, which can be adapted to various attention-related models, reveals that the current Transformer architecture has the potential for further evolution. Extensive experiments demonstrate that treating attention as a feature map and applying convolution as a processing method significantly enhances Transformer performance.

MuSC: Improving Complex Instruction Following with Multi-granularity Self-Contrastive Training

Hui Huang, Jiaheng Liu, Yancheng He, Shilong Li, Bing Xu, Conghui Zhu, Muyun Yang, Tiejun Zhao

Complex instruction-following with elaborate constraints is imperative for Large Language Models (LLMs). While existing methods have constructed data for complex instruction alignment, they all rely on a more advanced model, especially GPT-4, limiting their application. In this paper, we propose a Multi-granularity Self-Contrastive Training (MuSC) framework, to improve the complex instruction alignment without relying on a stronger model. Our method is conducted on both coarse and fine granularity. On coarse-granularity, we construct constraint-aware preference data based on instruction decomposition and recombination. On fine-granularity, we perform token-aware preference optimization with dynamic token-level supervision. Our method is evaluated on open-sourced models, and experiment results show our method achieves significant improvement on both complex and general instruction-following benchmarks, surpassing previous self-alignment methods.

Powerformer: Efficient and High-Accuracy Privacy-Preserving Language Model with Homomorphic Encryption

Dongjin Park, Eunsang Lee, Joon-Woo Lee

We propose Powerformer, an efficient homomorphic encryption (HE)-based privacy-preserving language model (PPLM) designed to reduce computation overhead while maintaining model performance. Powerformer incorporates three key techniques to optimize encrypted computations: 1. A novel distillation technique that replaces softmax and layer normalization (LN) with computationally efficient power and linear functions, ensuring no performance degradation while enabling seamless encrypted computation. 2. A pseudo-sign composite approximation method that accurately approximates GELU and tanh functions with minimal computational overhead. 3. A homomorphic matrix multiplication algorithm specifically optimized for Transformer models, enhancing efficiency in encrypted environments. By integrating these techniques, Powerformer based on the BERT-base model achieves a 45% reduction in computation time compared to the state-of-the-art HE-based PPLM without any loss in accuracy.

Scaling Laws for RNN LLM in Long-Context Scenarios with State Size

Kai Liu, Jianfei Gao, Kai Chen

The Transformer architecture has become the standard LLM architecture due to its powerful self-attention mechanism. However, it suffers from quadratic computational complexity and linear memory complexity. RNN-based LLMs have been proposed as alternatives. Yet, RNN models struggle in long-context scenarios, making it challenging to replace self-attention with RNNs. We identify the state size as a critical bottleneck, which is significantly smaller than that of Transformers with a basic context length of 2k. However, simply increasing the state size significantly raises the number of parameters and lowers training efficiency. In this paper, we propose an efficient scaling method to scale the state size of RNN models to match the 2k context length of Transformers, with small parameters overhead. Experimental results demonstrate that scaling the state size significantly enhances long-context understanding. Retrieval performance scales almost linearly with state size, with a 454M model featuring an expanded state achieving performance comparable to a 1.47B model on FDA, a recall-intensive task. These findings highlight state scaling as a promising approach for advancing RNN-based LLMs.

ACECODER: Acing Coder RL via Automated Test-Case Synthesis

Huaye Zeng, Dongfu Jiang, Haozhe Wang, Ping Nie, Xiaotong Chen, Wenhu Chen

Most progress in recent coder models has been driven by supervised fine-tuning (SFT), while the potential of reinforcement learning (RL) remains largely unexplored, primarily due to the lack of reliable reward data/model in the code domain. In this paper, we address this challenge by leveraging automated large-scale test-case synthesis to enhance code model training. Specifically, we design a pipeline that generates extensive (question, test-cases) pairs from existing code data. Using these test cases, we construct preference pairs based on pass rates over sampled programs to train reward models with Bradley-Terry loss. It shows an average of 10-point improvement for Llama-3.1-8B-Ins and 5-point improvement for Qwen2.5-Coder-7B-Ins through best-of-32 sampling, making the 7B model on par with 236B DeepSeek-V2.5. Furthermore, we conduct reinforcement learning with both reward models and test-case pass rewards, leading to consistent improvements across HumanEval, MBPP, BigCodeBench, and LiveCodeBench (V4). Notably, we follow the R1-style training to start from Qwen2.5-Coder-base directly and show that our RL training can improve model on HumanEval-plus by over 25% and MBPP-plus by 6% for merely 80 optimization steps. We believe our results highlight the huge potential of reinforcement learning in coder models.

Finding the Sweet Spot: Preference Data Construction for Scaling Preference Optimization

Yao Xiao, Hai Ye, Linyao Chen, Hwee Tou Ng, Lidong Bing, Xiaoli Li, Roy Ka-Wei Lee

Iterative data generation and model retraining are widely used to align large language models (LLMs). It typically involves a policy model to generate on-policy responses and a reward model to guide training data selection. Direct Preference Optimization (DPO) further enhances this process by constructing preference pairs of chosen and rejected responses. In this work, we aim to \emph{scale up} the number of on-policy samples via repeated random sampling to improve alignment performance. Conventional practice selects the sample with the highest reward as chosen and the lowest as rejected for DPO. However, our experiments reveal that this strategy leads to a \emph{decline} in performance as the sample size increases. To address this, we investigate preference data construction through the lens of underlying normal distribution of sample rewards. We categorize the reward space into seven representative points and systematically explore all 21 (C_7^2) pairwise combinations. Through evaluations on four models using AlpacaEval 2, we find that selecting the rejected response at reward position \mu - 2\sigma rather than the minimum reward, is crucial for optimal performance. We finally introduce a scalable preference data construction strategy that consistently enhances model performance as the sample scale increases.

HFT: Half Fine-Tuning for Large Language Models

Tingfeng Hui, Zhenyu Zhang, Shuohuan Wang, Weiran Xu, Yu Sun, Hua Wu

Large language models (LLMs) with one or more fine-tuning phases have become necessary to unlock various capabilities, enabling LLMs to follow natural language instructions and align with human preferences. However, it carries the risk of catastrophic forgetting during sequential training, the parametric knowledge or the ability learned in previous stages may be overwhelmed by incoming training data. This paper finds that LLMs can restore some original knowledge by regularly resetting partial parameters. Inspired by this, we introduce \underline{H}alf \underline{F}ine-\underline{T}uning (HFT) for LLMs, as a substitute for full fine-tuning (FFT), to mitigate the forgetting issues, where half of the parameters are selected to learn new tasks. In contrast, the other half are frozen to retain previous knowledge. We provide a feasibility analysis from the optimization perspective and interpret the parameter selection operation as a regularization term. HFT could be seamlessly integrated into existing fine-tuning frameworks without changing the model architecture. Extensive experiments and analysis on supervised fine-tuning, direct preference optimization, and continual learning consistently demonstrate the effectiveness, robustness, and efficiency of HFT. Compared with FFT, HFT not only significantly alleviates the forgetting problem, but also achieves the best performance in a series of downstream benchmarks, with an approximately 30% reduction in training time.

Upcycling Instruction Tuning from Dense to Mixture-of-Experts via Parameter Merging

Tingfeng Hui, Zhenyu Zhang, Shuohuan Wang, Yu Sun, Hua Wu, Sen Su

Mixture-of-Experts (MoE) shines brightly in large language models (LLMs) and demonstrates outstanding performance in plentiful natural language processing tasks. However, existing methods transforming LLMs from dense to MoE face significant data requirements and typically rely on large-scale post-training. In this paper, we propose Upcycling Instruction Tuning (UpIT), a data-efficient approach for tuning a dense pre-trained model into a MoE instruction model. Specifically, we first point out that intermediate checkpoints during instruction tuning of the dense model are naturally suitable for specialized experts, and then propose an expert expansion stage to flexibly achieve models with flexible numbers of experts, where genetic algorithm and parameter merging are introduced to ensure sufficient diversity of new extended experts. To ensure that each specialized expert in the MoE model works as expected, we select a small amount of seed data that each expert excels to pre-optimize the router. Extensive experiments with various data scales and upcycling settings demonstrate the outstanding performance and data efficiency of UpIT, as well as stable improvement in expert or data scaling. Further analysis reveals the importance of ensuring expert diversity in upcycling.

Code-Switching Red-Teaming: LLM Evaluation for Safety and Multilingual Understanding

Haneul Yoo, Yongjin Yang, Hwaran Lee

As large language models (LLMs) have advanced rapidly, concerns regarding their safety have become prominent. In this paper, we discover that code-switching in red-teaming queries can effectively elicit undesirable behaviors of LLMs, which are common practices in natural language. We introduce a simple yet effective framework, CSRT, to synthesize code-switching red-teaming queries and investigate the safety and multilingual understanding of LLMs comprehensively. Through extensive experiments with ten state-of-the-art LLMs and code-switching queries combining up to 10 languages, we demonstrate that the CSRT significantly outperforms existing multilingual red-teaming techniques, achieving 46.7% more attacks than standard attacks in English and being effective in conventional safety domains. We also examine the multilingual ability of those LLMs to generate and understand code-switching texts. Additionally, we validate the extensibility of the CSRT by generating code-switching attack prompts with monolingual data. We finally conduct detailed ablation studies exploring code-switching and propound unintended correlation between resource availability of languages and safety alignment in existing multilingual LLMs.

Mimicking the Familiar: Dynamic Command Generation for Information Theft Attacks in LLM Tool-Learning System

Ziyou Jiang, Mingyang Li, Guowei Yang, Junjie Wang, Yuekai Huang, Zhiyuan Chang, Qing Wang

Information theft attacks pose a significant risk to Large Language Model (LLM) tool-learning systems. Adversaries can inject malicious commands through compromised tools, manipulating LLMs to send sensitive information to these tools, which leads to potential privacy breaches. However, existing attack approaches are black-box oriented and rely on static commands that cannot adapt flexibly to the changes in user queries and the invocation chain of tools. It makes malicious commands more likely to be detected by LLM and leads to attack failure. In this paper, we propose AutoCMD, a dynamic attack comment generation approach for information theft attacks in LLM tool-learning systems. Inspired by the concept of mimicking the familiar, AutoCMD is capable of inferring the information utilized by upstream tools in the toolchain through learning on open-source systems and reinforcement with target system examples, thereby generating more targeted commands for information theft. The evaluation results show that AutoCMD outperforms the baselines with +13.2% ASR_{Theft}, and can be generalized to new tool-learning systems to expose their information leakage risks. We also design four defense methods to effectively protect tool-learning systems from the attack.

When GPT Spills the Tea: Comprehensive Assessment of Knowledge File Leakage in GPTs

Xinyue Shen, Yun Shen, Michael Backes, Yang Zhang

Knowledge files have been widely used in large language model (LLM)-powered agents, such as GPTs, to improve response quality. However, concerns over the potential leakage of knowledge files have grown significantly. Existing studies demonstrate that adversarial prompts can induce GPTs to leak knowledge file content. Yet, it remains uncertain whether additional leakage vectors exist, particularly given the complex data flows across clients, servers, and databases in GPTs. In this paper, we present a comprehensive risk assessment of knowledge file leakage, leveraging a novel workflow inspired by Data Security Posture Management (DSPM). Through the analysis of 651,022 GPT metadata, 11,820 flows, and 1,466 responses, we identify five leakage vectors: metadata, GPT initialization, retrieval, sandboxed execution environments, and prompts. These vectors enable adversaries to extract sensitive knowledge file data such as titles, content, types, and sizes. Notably, the activation of the built-in tool Code Interpreter leads to a privilege escalation vulnerability, enabling adversaries to directly download original knowledge files with a 95.95% success rate. Further analysis reveals that 28.80% of leaked files are copyrighted, including digital copies from major publishers and internal materials from a listed company. In the end, we provide actionable solutions for GPT builders and platform providers to secure the GPT data supply chain.

DIVE into MoE: Diversity-Enhanced Reconstruction of Large Language Models from Dense into Mixture-of-Experts

Yuchen Feng, Bowen Shen, Naibin Gu, Jiaxuan Zhao, Peng Fu, Zheng Lin, Weiping Wang

Large language models (LLMs) with the Mixture-of-Experts (MoE) architecture achieve high cost-efficiency by selectively activating a subset of the parameters. Despite the inference efficiency of MoE LLMs, the training of extensive experts from scratch incurs substantial overhead, whereas reconstructing a dense LLM into an MoE LLM significantly reduces the training budget. However, existing reconstruction methods often overlook the diversity among experts, leading to potential redundancy. In this paper, we come up with the observation that a specific LLM exhibits notable diversity after being pruned on different calibration datasets, based on which we present a Diversity-Enhanced reconstruction method named DIVE. The recipe of DIVE includes domain affinity mining, pruning-based expert reconstruction, and efficient retraining. Specifically, the reconstruction includes pruning and reassembly of the feed-forward network (FFN) module. After reconstruction, we efficiently retrain the model on routers, experts and normalization modules. We implement DIVE on Llama-style LLMs with open-source training corpora. Experiments show that DIVE achieves training efficiency with minimal accuracy trade-offs, outperforming existing pruning and MoE reconstruction methods with the same number of activated parameters. Code is available at: https://github.com/yuchenblah/DIVE.

AlignDistil: Token-Level Language Model Alignment as Adaptive Policy Distillation

Songming Zhang, Xue Zhang, Tong Zhang, Bojie Hu, Yufeng Chen, Jinan Xu

In modern large language models (LLMs), LLM alignment is of crucial importance and is typically achieved through methods such as reinforcement learning from human feedback (RLHF) and direct preference optimization (DPO). However, in most existing methods for LLM alignment, all tokens in the response are optimized using a sparse, response-level reward or preference annotation. The ignorance of token-level rewards may erroneously punish high-quality tokens or encourage low-quality tokens, resulting in suboptimal performance and slow convergence speed. To address this issue, we propose \emph{\textbf{AlignDistil}}, a RLHF-equivalent distillation method for token-level reward optimization. Specifically, we introduce the reward learned by DPO into the RLHF objective and theoretically prove the equivalence between this objective and a token-level distillation process, where the teacher distribution linearly combines the logits from the DPO model and a reference model. On this basis, we further bridge the accuracy gap between the reward from the DPO model and the pure reward model, by building a contrastive DPO reward with a normal and a reverse DPO model. Moreover, to avoid under- and over-optimization on different tokens, we design a token adaptive logit extrapolation mechanism to construct an appropriate teacher distribution for each token. Experimental results demonstrate the superiority of our AlignDistil over existing methods and showcase fast convergence due to its token-level distributional reward optimization.

Multi-Attribute Steering of Language Models via Targeted Intervention

Duy Nguyen, Archiki Prasad, Elias Stengel-Eskin, Mohit Bansal

Inference-time intervention (ITI) has emerged as a promising method for steering large language model (LLM) behavior in a particular direction (e.g., improving helpfulness) by intervening on token representations without costly updates to the LLM's parameters. However, existing ITI approaches fail to scale to multi-attribute settings with conflicts, such as enhancing helpfulness while also reducing toxicity. To address this, we introduce Multi-Attribute Targeted Steering (MAT-Steer), a novel steering framework designed for selective token-level intervention across multiple attributes. We achieve this by learning steering vectors using an alignment objective that shifts the model's internal representations of undesirable outputs closer to those of desirable ones while enforcing sparsity and orthogonality among vectors for different attributes, thereby reducing inter-attribute conflicts. We evaluate MAT-Steer in two distinct settings: (i) on question answering (QA) tasks where we balance attributes like truthfulness, bias, and toxicity; (ii) on generative tasks where we simultaneously improve attributes like helpfulness, correctness, and coherence. MAT-Steer outperforms existing ITI and parameter-efficient fine-tuning approaches across both task types (e.g., average 3% accuracy gain across QA tasks and 55.82% win rate against the best ITI baseline).

Optimal Transport-Based Token Weighting scheme for Enhanced Preference Optimization

Meng Li, Guangda Huzhang, Haibo Zhang, Xiting Wang, Anxiang Zeng

Direct Preference Optimization (DPO) has emerged as a promising framework for aligning Large Language Models (LLMs) with human preferences by directly optimizing the log-likelihood difference between chosen and rejected responses. However, existing methods assign equal importance to all tokens in the response, while humans focus on more meaningful parts. This leads to suboptimal preference optimization, as irrelevant or noisy tokens disproportionately influence DPO loss. To address this limitation, we propose \textbf{O}ptimal \textbf{T}ransport-based token weighting scheme for enhancing direct \textbf{P}reference \textbf{O}ptimization (OTPO). By emphasizing semantically meaningful token pairs and de-emphasizing less relevant ones, our method introduces a context-aware token weighting scheme that yields a more contrastive reward difference estimate. This adaptive weighting enhances reward stability, improves interpretability, and ensures that preference optimization focuses on meaningful differences between responses. Extensive experiments have validated OTPO's effectiveness in improving instruction-following ability across various settings\footnote{Code is available at https://github.com/Mimasss2/OTPO.}.

RiOT: Efficient Prompt Refinement with Residual Optimization Tree

Chenyi Zhou, Zhengyan Shi, Yuan Yao, Lei Liang, Huajun Chen, Qiang Zhang

Recent advancements in large language models (LLMs) have highlighted their potential across a variety of tasks, but their performance still heavily relies on the design of effective prompts. Existing methods for automatic prompt optimization face two challenges: lack of diversity, limiting the exploration of valuable and innovative directions and semantic drift, where optimizations for one task can degrade performance in others. To address these issues, we propose Residual Optimization Tree (RiOT), a novel framework for automatic prompt optimization. RiOT iteratively refines prompts through text gradients, generating multiple semantically diverse candidates at each step, and selects the best prompt using perplexity. Additionally, RiOT incorporates the text residual connection to mitigate semantic drift by selectively retaining beneficial content across optimization iterations. A tree structure efficiently manages the optimization process, ensuring scalability and flexibility. Extensive experiments across five benchmarks — covering commonsense, mathematical, logical, temporal, and semantic reasoning — demonstrate that RiOT outperforms both previous prompt optimization methods and manual prompting. Code will be released.

HoPE: A Novel Positional Encoding Without Long-Term Decay for Enhanced Context Awareness and Extrapolation

Yuhan Chen, Ang Lv, Jian Luan, Bin Wang, Wei Liu

Many positional encodings (PEs) are designed to exhibit long-term decay, based on an entrenched and long-standing inductive opinion: tokens farther away from the current position carry less relevant information. We argue that long-term decay is outdated in the era of LLMs, as LLMs are now applied to tasks demanding precise retrieval of in-context information from arbitrary positions. Firstly, we present empirical analyses on various PEs, demonstrating that models inherently learn attention with only a local-decay pattern while forming a U-shape pattern globally, contradicting the principle of long-term decay. Furthermore, we conduct a detailed analysis of rotary position encoding (RoPE, a prevalent relative positional encoding in LLMs), and found that the U-shape attention is caused by some learned components, which are also the key factor limiting RoPE's expressiveness and extrapolation. Inspired by these insights, we propose High-frequency rotary Position Encoding (HoPE). HoPE replaces the specific components in RoPE with position-independent ones, retaining only high-frequency signals, which also breaks the principle of long-term decay in theory. HoPE achieves two major advantages: (1) Without constraints imposed by long-term decay, contradictory factors that limit attention optimization are removed. Thus, the model's context awareness is enhanced. (2) HoPE exhibits greater robustness to the out-of-distribution behavior in attention patterns during extrapolation. The effectiveness of HoPE is validated through extensive experiments and with a large language model of up to 3 billion parameters.

Pre-Training Curriculum for Multi-Token Prediction in Language Models

Ansar Aynetdinov, Alan Akbik

Multi-token prediction (MTP) is a recently proposed pre-training objective for language models. Rather than predicting only the next token (NTP), MTP predicts the next *k* tokens at each prediction step, using multiple prediction heads. MTP has shown promise in improving downstream performance, inference speed, and training efficiency, particularly for large models. However, prior work has shown that smaller language models (SLMs) struggle with the MTP objective. To address this, we propose a curriculum learning strategy for MTP training, exploring two variants: a forward curriculum, which gradually increases the complexity of the pre-training objective from NTP to MTP, and a reverse curriculum, which does the opposite. Our experiments show that the forward curriculum enables SLMs to better leverage the MTP objective during pre-training, improving downstream NTP performance and generative output quality, while retaining the benefits of self-speculative decoding. The reverse curriculum achieves stronger NTP performance and output quality, but fails to provide any self-speculative decoding benefits.

Gödel Agent: A Self-Referential Agent Framework for Recursively Self-Improvement

Xunjian Yin, Xinyi Wang, Liangming Pan, Li Lin, Xiaojun Wan, William Yang Wang

The rapid advancement of large language models (LLMs) has significantly enhanced the capabilities of agents across various tasks. However, existing agentic systems, whether based on fixed pipeline algorithms or pre-defined meta-learning frameworks, cannot search the whole agent design space due to the restriction of human-designed components, and thus might miss the more optimal agent design. In this paper, we introduce Gödel Agent, a self-evolving framework inspired by the Gödel Machine, enabling agents to recursively improve themselves without relying on predefined routines or fixed optimization algorithms. Gödel Agent leverages LLMs to dynamically modify its own logic and behavior, guided solely by high-level objectives through prompting. Experimental results on multiple domains demonstrate that the implementation of Gödel Agent can achieve continuous self-improvement, surpassing manually crafted agents in performance, efficiency, and generalizability.

SGIC: A Self-Guided Iterative Calibration Framework for RAG

Guanhua Chen, Yutong Yao, Lidia S. Chao, Xuebo Liu, Derek F. Wong

Recent research in retrieval-augmented generation (RAG) has concentrated on retrieving useful information from candidate documents. However, numerous methodologies frequently neglect the calibration capabilities of large language models (LLMs), which capitalize on their robust in-context reasoning prowess. This work illustrates that providing LLMs with specific cues substantially improves their calibration efficacy, especially in multi-round calibrations. We present a new SGIC: Self-Guided Iterative Calibration Framework that employs uncertainty scores as a tool. Initially, this framework calculates uncertainty scores to determine both the relevance of each document to the query and the confidence level in the responses produced by the LLMs. Subsequently, it reevaluates these scores iteratively, amalgamating them with prior responses to refine calibration. Furthermore, we introduce an innovative approach for constructing an iterative self-calibration training set, which optimizes LLMs to efficiently harness uncertainty scores for capturing critical information and enhancing response accuracy. Our proposed framework significantly improves performance on both closed-source and open-source LLMs.

Agentic Reasoning: A Streamlined Framework for Enhancing LLM Reasoning with Agentic Tools

Junde Wu, Jiayuan Zhu, Yuyuan Liu, Min Xu, Yueming Jin

We introduce Agentic Reasoning, a framework that enhances large language model (LLM) reasoning by integrating external tool-using agents. Agentic Reasoning dynamically leverages web search, code execution, and structured memory to address complex problems requiring deep research. A key innovation in our framework is the Mind-Map agent, which constructs a structured knowledge graph to store reasoning context and track logical relationships, ensuring coherence in long reasoning chains with extensive tool usage. Additionally, we conduct a comprehensive exploration of the Web-Search agent, leading to a highly effective search mechanism that surpasses all prior approaches. When deployed on DeepSeek-R1, our method achieves a new state-of-the-art (SOTA) among public models and delivers performance comparable to OpenAI Deep Research, the leading proprietary model in this domain. Extensive ablation studies validate the optimal selection of agentic tools and confirm the effectiveness of our Mind-Map and Web-Search agents in enhancing LLM reasoning. Our code and data are publicly available.

MDCure: A Scalable Pipeline for Multi-Document Instruction-Following

Gabrielle Kaili-May Liu, Bowen Shi, Avi Caciularu, Idan Szpektor, Arman Cohan

Multi-document (MD) processing is crucial for LLMs to handle real-world tasks such as summarization and question-answering across large sets of documents. While LLMs have improved at processing long inputs, MD contexts still present unique difficulties, including management of inter-document dependencies, redundancy, and incoherent structures. To address this challenge, we introduce MDCure, a scalable and effective instruction data generation framework to enhance the MD capabilities of LLMs without the computational cost of pre-training or reliance on human-annotated data. MDCure generates high-quality synthetic MD instruction data over sets of articles via targeted prompts. We also introduce MDCureRM, a cost-effective, MD-specific reward model to score and filter generated data based on their training utility for MD settings. MDCure is compatible with open- and closed-source models in addition to policy optimization methods such as PPO, enabling even small open- source models to surpass proprietary LLMs as strong generators of high-quality MD instruction data without further data filtering. With MDCure, we fine-tune a wide variety of LLMs up to 70B parameters in size from the FlanT5, Qwen2, and LLAMA3.1 model families. Extensive evaluations on a wide range of MD and long-context benchmarks spanning various tasks and domains show MDCure consistently improves performance over pre-trained baselines and base models by up to 75.1%.

More is not always better? Enhancing Many-Shot In-Context Learning with Differentiated and Reweighting Objectives

Xiaoqing Zhang, Ang Lv, Yuhan Liu, Flood Sung, Wei Liu, Jian Luan, Shuo Shang, Xiuying Chen, Rui Yan

Large language models (LLMs) excel at few-shot in-context learning (ICL) without requiring parameter updates. However, as ICL demonstrations increase from a few to many, performance tends to plateau and eventually decline. We identify two primary causes for this trend: the suboptimal negative log-likelihood (NLL) optimization objective and the incremental data noise. To address these issues, we introduce \textit{DrICL}, a novel optimization method that enhances model performance through \textit{Differentiated} and \textit{Reweighting} objectives. Globally, DrICL utilizes differentiated learning to optimize the NLL objective, ensuring that many-shot performance surpasses zero-shot levels. Locally, it dynamically adjusts the weighting of many-shot demonstrations by leveraging cumulative advantages inspired by reinforcement learning, thereby mitigating the impact of noisy data. Recognizing the lack of multi-task datasets with diverse many-shot distributions, we develop the \textit{Many-Shot ICL Benchmark} (ICL-50)-a large-scale benchmark of 50 tasks that cover shot numbers from 1 to 350 within sequences of up to 8,000 tokens-for both fine-tuning and evaluation purposes. Experimental results demonstrate that LLMs enhanced with DrICL achieve significant improvements in many-shot setups across various tasks, including both in-domain and out-of-domain scenarios. We release the code and dataset hoping to facilitate further research in many-shot ICL\footnote{https://github.com/xiaoqzhwhu/DrICL}.

FocusLLM: Precise Understanding of Long Context by Dynamic Condensing

Zhenyu Li, Yike Zhang, Tengyu Pan, Yutao Sun, Zhichao Duan, Junjie Fang, Rong Han, Zixuan Wang, Jianyong Wang

Empowering LLMs with the ability to precisely understand long contexts is crucial for many downstream applications. However, handling long contexts with conventional transformer architecture requires substantial training and inference resources. Existing context condensing methods cannot accurately understand the full context, as there is a considerable amount of information loss in the condensing process. To address these issues, we present **FocusLLM**, a framework designed to extend the fixed context length of any decoder-only LLM, allowing the model to focus on relevant information from very long sequences. FocusLLM first divides long text input into chunks based on the model's original context length. It then employs the **_dynamic condensing_** process to distill crucial information from each chunk. Ultimately, through the novel **_parallel decoding_** mechanism, FocusLLM can integrate the extracted information into its local context. FocusLLM stands out for great training efficiency and versatility: trained with an 8K input length and with much less training cost than previous methods, FocusLLM exhibits superior performance across downstream tasks and maintains strong language modeling ability when handling extensive long texts, even up to 400K tokens. Our code is available at https://github.com/leezythu/FocusLLM.

Optimizing Pre-Training Data Mixtures with Mixtures of Data Expert Models

Lior Belenki, Alekh Agarwal, Tianze Shi, Kristina Toutanova

We propose a method to optimize language model pre-training data mixtures through efficient approximation of the cross-entropy loss corresponding to each candidate mixture via a Mixture of Data Experts (MDE). We use this approximation as a source of additional features in a regression model, trained from observations of model loss for a small number of mixtures. Experiments with Transformer decoder-only language models in the range of 70M to 10B parameters on the SlimPajama dataset show that our method achieves significantly better performance than approaches that train regression models using only the mixture rates as input features. Combining this improved optimization method with an objective that takes into account cross-entropy on end task data leads to superior performance on few-shot downstream evaluations. We also provide theoretical insights on why aggregation of data expert predictions can provide good approximations to model losses for data mixtures.

Magnet: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation

Fan Yin, Zifeng Wang, I-Hung Hsu, Jun Yan, Ke Jiang, Yanfei Chen, Jindong Gu, Long Le, Kai-Wei Chang, Chen-Yu Lee, Hamid Palangi, Tomas Pfister

Large language models (LLMs) have exhibited the ability to effectively utilize external tools to address user queries. However, their performance may be limited in complex, multi-turn interactions involving users and multiple tools. To address this, we propose \textsc{Magnet}, a principled framework for synthesizing high-quality training trajectories to enhance the function calling capability of large language model agents in multi-turn conversations with humans. The framework is based on automatic and iterative translations from a function signature path to a sequence of queries and executable function calls. We model the complicated function interactions in multi-turn cases with graph and design novel node operations to build reliable signature paths. Motivated by context distillation, when guiding the generation of positive and negative trajectories using a teacher model, we provide reference function call sequences as positive hints in context and contrastive, incorrect function calls as negative hints. Experiments show that training with the positive trajectories with supervised fine-tuning and preference optimization against negative trajectories, our 14B model, \textsc{Magnet}-14B-mDPO, obtains 68.01 on BFCL-v3 and 73.30 on ToolQuery, surpassing the performance of the teacher model Gemini-1.5-pro-002 by a large margin in function calling.

Just Go Parallel: Improving the Multilingual Capabilities of Large Language Models

Muhammad Reza Qorib, Junyi Li, Hwee Tou Ng

Large language models (LLMs) have demonstrated impressive translation capabilities even without being explicitly trained on parallel data. This remarkable property has led some to believe that parallel data is no longer necessary for building multilingual language models. While some attribute this to the emergent abilities of LLMs due to scale, recent work suggests that it is actually caused by incidental bilingual signals present in the training data. Various methods have been proposed to maximize the utility of parallel data to enhance the multilingual capabilities of multilingual encoder-based and encoder-decoder language models. However, some decoder-based LLMs opt to ignore parallel data instead. In this work, we conduct a systematic study on the impact of adding parallel data on LLMs' multilingual capabilities, focusing specifically on translation and multilingual common-sense reasoning. Through controlled experiments, we demonstrate that parallel data can significantly improve LLMs' multilingual capabilities.

Session 12: Linguistic theories, Cognitive Modeling and Psycholinguistics (Wednesday 30 July, 11:00-12:30, Hall 4/5)

PsyDT: Using LLMs to Construct the Digital Twin of Psychological Counselor with Personalized Counseling Style for Psychological Counseling

Haojie Xie, Yirong Chen, Xiaofen Xing, Jingkai Lin, Xiangmin Xu

Currently, large language models (LLMs) have made significant progress in the field of psychological counseling. However, existing mental health LLMs overlook a critical issue where they do not consider the fact that different psychological counselors exhibit different personal styles, including linguistic style and therapy techniques, etc. As a result, these LLMs fail to satisfy the individual needs of clients who seek different counseling styles. To help bridge this gap, we propose PsyDT, a novel framework using LLMs to construct the Digital Twin of Psychological counselor with personalized counseling style. Compared to the time-consuming and costly approach of collecting a large number of real-world counseling cases to create a specific counselor's digital twin, our framework offers a faster and more cost-effective solution. To construct PsyDT, we utilize dynamic one-shot learning by using GPT-4 to capture counselor's unique counseling style, mainly focusing on linguistic style and therapy techniques. Subsequently, using existing single-turn long-text dialogues with client's questions, GPT-4 is guided to synthesize multi-turn dialogues of specific counselor. Finally, we fine-tune the LLMs on the synthetic dataset, PsyDTCorpus, to achieve the digital twin of psychological counselor with personalized counseling style. Experimental results indicate that our proposed PsyDT framework can synthesize multi-turn dialogues that closely resemble real-world counseling cases and demonstrate better performance compared to other baselines, thereby show that our framework can effectively construct the digital twin of psychological counselor with a specific counseling style.

Automatic detection of dyslexia based on eye movements during reading in Russian

Anna Laurinavichyute, Anastasiya Lopukhina, David Robert Reich

Dyslexia, a common learning disability, requires an early diagnosis. However, current screening tests are very time- and resource-consuming. We present an LSTM that aims to automatically classify dyslexia based on eye movements recorded during natural reading combined with basic demographic information and linguistic features. The proposed model reaches an AUC of 0.93 and outperforms the state-of-the-art model by 7 %. We report several ablation studies demonstrating that the fixation features matter the most for classification.

Comparing Moral Values in Western English-speaking societies and LLMs with Word Associations

Chaoyi Xiang, Chunhua Liu, Simon De Deyne, Lea Frermann

As the impact of large language models increases, understanding the moral values they encode becomes ever more important. Assessing moral values encoded in these models via direct prompting is challenging due to potential leakage of human norms into model training data, and their sensitivity to prompt formulation. Instead, we propose to use word associations, which have been shown to reflect moral reasoning in humans, as low-level underlying representations to obtain a more robust picture of LLMs' moral reasoning. We study moral differences in associations from western English-speaking communities and LLMs trained predominantly on English data. First, we create a large dataset of \textit{LLM-generated} word associations, resembling an existing data set of \textit{human} word associations. Next, we propose a novel method to propagate moral values based on seed words derived from Moral Foundation Theory through the human and LLM-generated association graphs. Finally, we compare the resulting moral representations, highlighting detailed but systematic differences between moral values emerging from English speakers and LLM associations.

Conformity in Large Language Models

Xiaochen Zhu, Caiqi Zhang, Tom Stafford, Nigel Collier, Andreas Vlachos

The conformity effect describes the tendency of individuals to align their responses with the majority. Studying this bias in large language models (LLMs) is crucial, as LLMs are increasingly used in various information-seeking and decision-making tasks as conversation partners to improve productivity. Thus, conformity to incorrect responses can compromise their effectiveness. In this paper, we adapt psychological experiments to examine the extent of conformity in state-of-the-art LLMs. Our findings reveal that all models tested exhibit varying levels of conformity toward the majority, regardless of their initial choice or correctness, across different knowledge domains. Notably, we are the first to show that LLMs are more likely to conform when they are more uncertain in their own prediction. We further explore factors that influence conformity, such as training paradigms and input characteristics, finding that instruction-tuned models are less susceptible to conformity, while increasing the naturalness of majority tones amplifies conformity. Finally, we propose two interventions—Devil's Advocate and Question Distillation—to mitigate conformity, providing insights into building more robust language models.

Decoding Reading Goals from Eye Movements

Omer Shubi, Cfir Avraham Hadar, Yevgeni Berzak

Readers can have different goals with respect to the text that they are reading. Can these goals be decoded from their eye movements over the text? In this work, we examine for the first time whether it is possible to distinguish between two types of common reading goals: information seeking and ordinary reading for comprehension. Using large-scale eye tracking data, we address this task with a wide range of models that cover different architectural and data representation strategies, and further introduce a new model ensemble. We find that transformer-based models with scanpath representations coupled with language modeling solve it most successfully, and that accurate predictions can be made in real time, shortly after the participant started reading the text. We further introduce a new method for model performance analysis based on mixed effect modeling. Combining this method with rich textual annotations reveals key properties of textual items and participants that contribute to the difficulty of the task, and improves our understanding of the variability in eye movement patterns across the two reading regimes.

Uncovering Visual-Semantic Psycholinguistic Properties from the Distributional Structure of Text Embedding Space

Si Wu, Sebastian Bruch

Imageability (potential of text to evoke a mental image) and concreteness (perceptibility of text) are two psycholinguistic properties that link visual and semantic spaces. It is little surprise that computational methods that estimate them do so using parallel visual and semantic spaces, such as collections of image-caption pairs or multi-modal models. In this paper, we work on the supposition that text itself in an image-caption dataset offers sufficient signals to accurately estimate these properties. We hypothesize, in particular, that the peakedness of the neighborhood of a word in the semantic embedding space reflects its degree of imageability and concreteness. We then propose an unsupervised, distribution-free measure, which we call Neighborhood Stability Measure (NSM), that quantifies the sharpness of peaks. Extensive experiments show that NSM correlates more strongly with ground-truth ratings than existing unsupervised methods, and is a strong predictor of these properties for classification. Our code and data are available on GitHub (https://github.com/Artificial-Memory-Lab/imageability).

The Cross-linguistic Role of Animacy in Grammar Structures

Nina Gregorio, Matteo Gay, Sharon Goldwater, Edoardo Ponti

Animacy is a semantic feature of nominals and follows a hierarchy: personal pronouns > human > animate > inanimate. In several languages, animacy imposes hard constraints on grammar. While it has been argued that these constraints may emerge from universal soft tendencies, it has been difficult to provide empirical evidence for this conjecture due to the lack of data annotated with animacy classes. In this work, we first propose a method to reliably classify animacy classes of nominals in 11 languages from 5 families, leveraging multilingual large language models (LLMs) and word sense disambiguation datasets. Then, through this newly acquired data, we verify that animacy displays consistent cross-linguistic tendencies in terms of preferred morphosyntactic constructions, although not always in line with received wisdom: animacy in nouns correlates with the alignment role of agent, early positions in a clause, and syntactic pivot (e.g., for relativisation), but not necessarily with grammatical subjecthood. Furthermore, the behaviour of personal pronouns in the hierarchy is idiosyncratic as they are rarely plural and relativised, contrary to high-animacy nouns.

Subword models struggle with word learning, but surprisal hides it

Bastian Bunzeck, Sina Zarrieß

We study word learning in subword and character language models with the psycholinguistic lexical decision task. While subword LMs struggle to discern words and non-words with high accuracy, character LMs solve this task easily and consistently. Only when supplied with further contexts do subword LMs perform similarly to character models. Additionally, when looking at word-level and syntactic learning trajectories, we find that both processes are separable in character LMs. Word learning happens before syntactic learning, whereas both occur simultaneously in subword LMs. This raises questions about the adequacy of subword LMs for modeling language acquisition and positions character LMs as a viable alternative to study processes below the syntactic level.

Developmentally-plausible Working Memory Shapes a Critical Period for Language Acquisition

Masato Mita, Ryo Yoshida, Yohei Oseki

Large language models possess general linguistic abilities but acquire language less efficiently than humans. This study proposes a method for integrating the developmental characteristics of working memory during the critical period, a stage when human language acquisition is particularly efficient, into the training process of language models. The proposed method introduces a mechanism that initially constrains working memory during the early stages of training and gradually relaxes this constraint in an exponential manner as learning progresses. Targeted syntactic evaluation shows that the proposed method outperforms conventional methods without memory constraints or with static memory constraints. These findings not only provide new directions for designing data-efficient language models but also offer indirect evidence supporting the role of the developmental characteristics of working memory as the underlying mechanism of the critical period in language acquisition.

If Attention Serves as a Cognitive Model of Human Memory Retrieval, What is the Plausible Memory Representation?

Ryo Yoshida, Shinnosuke Isono, Kohei Kajikawa, Taiga Someya, Yushi Sugimoto, Yohei Oseki

Recent work in computational psycholinguistics has revealed intriguing parallels between attention mechanisms and human memory retrieval, focusing primarily on vanilla Transformers that operate on token-level representations. However, computational psycholinguistic research has also established that syntactic structures provide compelling explanations for human sentence processing that token-level factors cannot fully account for. In this paper, we investigate whether the attention mechanism of Transformer Grammar (TG), which uniquely operates on syntactic structures as representational units, can serve as a cognitive model of human memory retrieval, using Normalized Attention Entropy (NAE) as a linking hypothesis between models and humans. Our experiments demonstrate that TG's attention achieves superior predictive power for self-paced reading times compared to vanilla Transformer's, with further analyses revealing independent contributions from both models. These findings suggest that human sentence processing involves dual memory representations---one based on syntactic structures and another on token sequences---with attention serving as the general memory retrieval algorithm, while highlighting the importance of incorporating syntactic structures as representational units.

Mitigating Gender Confounding Bias from Spoken Language in Dementia Detection via Weight Masking

Zhecheng Sheng, Xiruo Ding, Brian Hur, Changye Li, Trevor Cohen, Serguei V. S. Pakhomov

Deep transformer models have been used to detect linguistic anomalies in patient transcripts for early Alzheimer’s disease (AD) screening. While pre-trained neural language models (LMs) fine-tuned on AD transcripts perform well, little research has explored the effects of the gender of the speakers represented by these transcripts. This work addresses gender confounding in dementia detection and proposes two methods: the Extended Confounding Filter and the Dual Filter, which isolate and ablate weights associated with gender. We evaluate these methods on dementia datasets with first-person narratives from patients with cognitive impairment and healthy controls. Our results show transformer models tend to overfit to training data distributions. Disrupting gender-related weights results in a deconfounded dementia classifier, with the trade-off of slightly reduced dementia detection performance.

A New Formulation of Zipf’s Meaning-Frequency Law through Contextual Diversity

Ryo Nagata, Kumiko Tanaka-Ishii

This paper proposes formulating Zipf's meaning-frequency law, the power law between word frequency and the number of meanings, as a relationship between word frequency and contextual diversity. The proposed formulation quantifies meaning counts as contextual diversity, which is based on the directions of contextualized word vectors obtained from a Language Model (LM). This formulation gives a new interpretation to the law and also enables us to examine it for a wider variety of words and corpora than previous studies have explored. In addition, this paper shows that the law becomes unobservable when the size of the LM used is small and that autoregressive LMs require much more parameters than masked LMs to be able to observe the law.

Meta-Learning Neural Mechanisms rather than Bayesian Priors

Michael Eric Goodale, Salvador Mascarenhas, Yair Lakretz

Children acquire language despite being exposed to several orders of magnitude less data than large language models require. Meta-learning has been proposed as a way to integrate human-like learning biases into neural-network architectures, combining both the structured generalizations of symbolic models with the scalability of neural-network models. But what does meta-learning exactly imbue the model with? We investigate the meta-learning of formal languages and find that, contrary to previous claims, meta-trained models are not learning simplicity-based priors when meta-trained on datasets organised around simplicity. Rather, we find evidence that meta-training imprints neural mechanisms (such as counters) into the model, which function like cognitive primitives for the network on downstream tasks. Most surprisingly, we find that meta-training on a *single* formal language can provide as much improvement to a model as meta-training on 5000 different formal languages, provided that the formal language incentivizes the learning of useful neural mechanisms. Taken together, our findings provide practical implications for efficient meta-learning paradigms and new theoretical insights into linking symbolic theories and neural mechanisms.

BIG5-CHAT: Shaping LLM Personalities Through Training on Human-Grounded Data

Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona T. Diab, Maarten Sap

In this work, we tackle the challenge of embedding realistic human personality traits into LLMs. Previous approaches have primarily focused on prompt-based methods that describe the behavior associated with the desired personality traits, suffering from realism and validity issues. To address these limitations, we introduce BIG5-CHAT, a large-scale dataset containing 100,000 dialogues designed to ground models in how humans express their personality in text. Leveraging this dataset, we explore Supervised Fine-Tuning and Direct Preference Optimization as training-based methods to align LLMs more naturally with human personality patterns. Our methods outperform prompting on personality assessments such as BFI and IPIP-NEO, with trait correlations more closely matching human data. Furthermore, our experiments reveal that models trained to exhibit higher conscientiousness, higher agreeableness, lower extraversion, and lower neuroticism display better performance on reasoning tasks, aligning with psychological findings on how these traits impact human cognitive performance. To our knowledge, this work is the first comprehensive study to demonstrate how training-based methods can shape LLM personalities through learning from real human behaviors.

On the Acquisition of Shared Grammatical Representations in Bilingual Language Models

Catherine Arnett, Tyler A. Chang, James A. Michaelov, Ben Bergen

Crosslingual transfer is crucial to contemporary language models' multilingual capabilities, but how it occurs is not well understood. We ask what happens to a monolingual language model when it begins to be trained on a second language. Specifically, we train small bilingual models for which we control the amount of data for each language and the order of language exposure. To find evidence of shared multilingual representations, we turn to structural priming, a method used to study grammatical representations in humans. We first replicate previous crosslingual structural priming results and find that after controlling for training data quantity and language exposure, there are asymmetrical effects across language pairs and directions. We argue that this asymmetry may shape hypotheses about human structural priming effects. We also find that structural priming effects are less robust for less similar language pairs, highlighting potential limitations of crosslingual transfer learning and shared representations for typologically diverse languages.

The Role of Abstract Representations and Observed Preferences in the Ordering of Binomials in Large Language Models

Zachary Nicholas Houghton, Kenji Sagae, Emily Morgan

To what extent do large language models learn abstract representations as opposed to more superficial aspects of their very large training corpora? We examine this question in the context of binomial ordering preferences involving two conjoined nouns in English. When choosing a binomial ordering (radio and television vs television and radio), humans rely on more than simply the observed frequency of each option. Humans also rely on abstract ordering preferences (e.g., preferences for short words before long words). We investigate whether large language models simply rely on the observed preference in their training data, or whether they are capable of learning the abstract ordering preferences (i.e., abstract representations) that humans rely on. Our results suggest that both smaller and larger models' ordering preferences are driven exclusively by their experience with that item in the training data. Our study provides further insights into differences between how large language models represent and use language and how humans do it, particularly with respect to the use of abstract representations versus observed preferences.

Language Models Grow Less Humanlike beyond Phase Transition

Tatsuya Aoyama, Ethan Wilcox

LMs' alignment with human reading behavior (i.e. psychometric predictive power; PPP) is known to improve during pretraining up to a tipping point, beyond which it either plateaus or degrades. Various factors, such as word frequency, recency bias in attention, and context size, have been theorized to affect PPP, yet there is no current account that explains why such a tipping point exists, and how it interacts with LMs' pretraining dynamics more generally. We hypothesize that the underlying factor is a pretraining phase transition, characterized by the rapid emergence of specialized attention heads. We conduct a series of correlational and causal experiments to show that such a phase transition is responsible for the tipping point in PPP. We then show that, rather than producing attention patterns that contribute to the degradation in PPP, phase transitions alter the subsequent learning dynamics of the model, such that further training keeps damaging PPP.

Coordinating Chaos: A Structured Review of Linguistic Coordination Methodologies

Benjamin Roger Litterer, David Jurgens, Dallas Card

Linguistic coordination—a phenomenon where conversation partners end up having similar patterns of language use—has been established across a variety of contexts and for multiple linguistic features. However, the study of language coordination has been accompanied by a diverse and inconsistently applied set of measures and theoretical perspectives. This diversity has significant consequences, as replication studies have highlighted the brittleness of certain measures and called influential findings into question. While prior work has addressed specific modeling decisions and model types, linguistic coordination research has yet to fully examine, synthesize, and critique the space of modeling choices available. In this work, we present a framework to organize the linguistic coordination literature. Using this schema, we provide a high-level overview of the choices involved in the measurement process and synthesize relevant critiques. Based on both gaps and limitations surfaced from this review, we suggest directions for further exploration and evaluation. In doing so, we provide the clarity required for linguistic coordination research to arrive at interpretable and sound conclusions.

LLMs syntactically adapt their language use to their conversational partner

Florian Kandra, Vera Demberg, Alexander Koller

It has been frequently observed that human speakers align their language use with each other during conversations. In this paper, we study empirically whether large language models (LLMs) exhibit the same behavior of conversational adaptation. We construct a corpus of conversations between LLMs and find that two LLM agents end up making more similar syntactic choices as conversations go on, confirming that modern LLMs adapt their language use to their conversational partners in at least a rudimentary way.

Anything Goes? A Crosslinguistic Study of (Im)possible Language Learning in LMs

Xiulin Yang, Tatsuya Aoyama, Yuekun Yao, Ethan Wilcox

Do language models (LMs) offer insights into human language learning? A common argument against this idea is that because their architecture and training paradigm are so vastly different from humans, LMs can learn arbitrary inputs as easily as natural languages. We test this claim by training LMs to model impossible and typologically unattested languages. Unlike previous work, which has focused exclusively on English, we conduct experiments on 12 languages from 4 language families with two newly constructed parallel corpora. Our results show that while GPT-2 small can largely distinguish attested languages from their impossible counterparts, it does not achieve perfect separation between all the attested languages and all the impossible ones. We further test whether GPT-2 small distinguishes typologically attested from unattested languages with different NP orders by manipulating word order based on Greenberg's Universal 20. We find that the model's perplexity scores do not distinguish attested vs. unattested word orders, while its performance on the generalization test does. These findings suggest that LMs exhibit some human-like inductive biases, though these biases are weaker than those found in human learners.

How LLMs Comprehend Temporal Structure in Narratives: A Case Study in Cognitive Evaluation of LLMs

Karin De Langis, Jong Inn Park, Andreas Schramm, Bin Hu, Khanh Chi Le, Dongyeop Kang

Large language models (LLMs) exihibit increasingly sophisticated linguistic capabilities, yet the extent to which these behaviors reflect human-like cognition versus advanced pattern recognition remains an open question. In this study, we investigate how LLMs process the temporal meaning of linguistic aspect in narratives that were previously used in human studies. Using an Expert-in-the-Loop probing pipeline, we conduct a series of targeted experiments to assess whether LLMs construct semantic representations and pragmatic inferences in a human-like manner. Our findings show that LLMs over-rely on prototypicality, produce inconsistent aspectual judgments, and struggle with causal reasoning derived from aspect, raising concerns about their ability to fully comprehend narratives. These results suggest that LLMs process aspect fundamentally differently from humans and lack robust narrative understanding. Beyond these empirical findings, we develop a standardized experimental framework for the reliable assessment of LLMs' cognitive and linguistic capabilities.

ScanEZ: Integrating Cognitive Models with Self-Supervised Learning for Spatiotemporal Scanpath Prediction

Ekta Sood, Prajit Dhar, Enrica Troiano, Rosy Southwell, Sidney K. DMello

Accurately predicting human scanpaths during reading is vital for diverse fields and downstream tasks, from educational technologies to automatic question answering. To date, however, progress in this direction remains limited by scarce gaze data. We overcome the issue with ScanEZ, a self-supervised framework grounded in cognitive models of reading. ScanEZ jointly models the spatial and temporal dimensions of scanpaths by leveraging synthetic data and a 3-D gaze objective inspired by masked language modeling. With this framework, we provide evidence that two key factors in scanpath prediction during reading are: the use of masked modeling of both spatial and temporal patterns of eye movements, and cognitive model simulations as an inductive bias to kick-start training. Our approach achieves state-of-the-art results on established datasets (e.g., up to 31.4% negative log-likelihood improvement on CELER L1), and proves portable across different experimental conditions.

The time scale of redundancy between prosody and linguistic context

Tamar I Regev, Chiebuka Ohams, Shaylee Xie, Lukas Wolf, Evelina Fedorenko, Alex Warstadt, Ethan Wilcox, Tiago Pimentel

In spoken communication, information is transmitted not only via words, but also through a rich array of non-verbal signals, including prosody---the non-segmental auditory features of speech. Do these different communication channels carry distinct information? Prior work has shown that the information carried by prosodic features is substantially redundant with that carried by the surrounding words. Here, we systematically examine the time scale of this relationship, studying how it varies with the length of past and future contexts. We find that a word's prosodic features require an extended past context (3-8 words across different features) to be reliably predicted. Given that long-scale contextual information decays in memory, prosody may facilitate communication by adding information that is locally unique. We also find that a word's prosodic features show some redundancy with future words, but only with a short scale of 1-2 words, consistent with reports of incremental short-term planning in language production. Thus, prosody may facilitate communication by helping listeners predict upcoming material. In tandem, our results highlight potentially distinct roles that prosody plays in facilitating integration of words into past contexts and in helping predict upcoming words.

Fine-Grained Spatio-Temporal Modeling of Reading Behavior

Francesco Ignazio Re, Andreas Opedal, Glib Manaiev, Mario Giulianelli, Ryan Cotterell

Reading is a process that unfolds across space and time. Standard modeling approaches, however, overlook much of the spatio-temporal dynamics involved in reading by relying on aggregated reading measurements---typically only focusing on fixation durations---and employing models with strong simplifying assumptions. In this paper, we propose a generative model that captures not only how long fixations last, but also where they land and when they occur. To this end, we model reading scanpaths via two conditionally independent distributions: one for fixation location and timing, and another for fixation duration. The location (and timing) of fixation shifts, so-called saccades, are modeled using a spatio-temporal Hawkes process, which captures how each fixation excites the probability of a new fixation occurring near it in time and space. Empirically, our Hawkes process model exhibits higher likelihood on held-out reading data than baselines. The duration time of fixation events is modeled as a function of fixation-specific features convolved across time, thus capturing non-stationary delayed effects. We find that convolution-based approaches demonstrate weak predictive power when modeling disaggregated fixation durations. Similarly, our analysis of surprisal theory on disaggregated data reveals limited effectiveness in predicting both where fixations occur and how long they last.

SocialEval: Evaluating Social Intelligence of Large Language Models

Jinfeng Zhou, Yuxuan Chen, Yihan Shi, Xuanming Zhang, Leqi Lei, Yi Feng, Zexuan Xiong, Miao Yan, Xunzhi Wang, Yaru Cao, Jianing Yin, Shuai Wang, Quanyu Dai, Zhenhua Dong, Hongning Wang, Minlie Huang

LLMs exhibit promising Social Intelligence (SI) in modeling human behavior, raising the need to evaluate LLMs' SI and their discrepancy with humans. SI equips humans with interpersonal abilities to behave wisely in navigating social interactions to achieve social goals. This presents an operational evaluation paradigm: outcome-oriented goal achievement evaluation and process-oriented interpersonal ability evaluation, which existing work fails to address. To this end, we propose SocialEval, a script-based bilingual SI benchmark, integrating outcome- and process-oriented evaluation by manually crafting narrative scripts. Each script is structured as a world tree that contains plot lines driven by interpersonal ability, providing a comprehensive view of how LLMs navigate social interactions. Experiments show that LLMs fall behind humans on both SI evaluations, exhibit prosociality, and prefer more positive social behaviors, even if they lead to goal failure. Analysis of LLMs' formed representation space and neuronal activations reveals that LLMs have developed ability-specific functional partitions akin to the human brain.

A Review of Theory of Mind Capabilities in Large Language Models

Ruirui Chen, Weifeng Jiang, Chengwei Qin, Cheston Tan

Theory of Mind (ToM)—the ability to reason about the mental states of oneself and others—is a cornerstone of human social intelligence. As Large Language Models (LLMs) become increasingly integrated into daily life, understanding their ability to interpret and respond to human mental states is crucial for enabling effective interactions. In this paper, we review LLMs' ToM capabilities by analyzing both evaluation benchmarks and enhancement strategies. For evaluation, we focus on recently proposed and widely used story-based benchmarks. For enhancement, we provide an in-depth analysis of recent methods aimed at improving LLMs' ToM abilities. Furthermore, we outline promising directions for future research to further advance these capabilities and better adapt LLMs to more realistic and diverse scenarios. Our survey serves as a valuable resource for researchers interested in evaluating and advancing LLMs' ToM capabilities.

The Harmonic Structure of Information Contours

Eleftheria Tsipidi, Samuel Kiegeland, Franz Nowak, Tianyang Xu, Ethan Wilcox, Alex Warstadt, Ryan Cotterell, Mario Giulianelli

The uniform information density (UID) hypothesis proposes that speakers aim to distribute information evenly throughout a text, balancing production effort and listener comprehension difficulty. However, language typically does not maintain a strictly uniform information rate; instead, it fluctuates around a global average. These fluctuations are often explained by factors such as syntactic constraints, stylistic choices, or audience design. In this work, we explore an alternative perspective: that these fluctuations may be influenced by an implicit linguistic pressure towards periodicity, where the information rate oscillates at regular intervals, potentially across multiple frequencies simultaneously. We apply harmonic regression and introduce a novel extension called time scaling to detect and test for such periodicity in information contours. Analyzing texts in English, Spanish, German, Dutch, Basque, and Brazilian Portuguese, we find consistent evidence of periodic patterns in information rate. Many dominant frequencies align with discourse structure, suggesting these oscillations reflect meaningful linguistic organization. Beyond highlighting the connection between information rate and discourse structure, our approach offers a general framework for uncovering structural pressures at various levels of linguistic granularity.

Session 12: NLP Applications (Wednesday 30 July, 11:00-12:30, Hall 4/5)

From Evasion to Concealment: Stealthy Knowledge Unlearning for LLMs

Tianle Gu, Kexin Huang, Ruilin Luo, Yuanqi Yao, Xiuying Chen, Yujiu Yang, Yan Teng, Yingchun Wang

LLM Unlearning plays a crucial role in removing sensitive information from language models to mitigate potential misuse. However, previous approaches often treat nonsensical responses or template-based refusals (e.g., ``Sorry, I cannot answer.") as the unlearning target, which can give the impression of deliberate information suppression, making the process even more vulnerable to attacks and jailbreaks. Moreover, most methods rely on auxiliary models or retaining datasets, which adds complexity to the unlearning process. To address these challenges, we propose MEOW, a streamlined and stealthy unlearning method that eliminates the need for auxiliary models or retaining data while avoiding leakage through its innovative use of inverted facts. These inverted facts are generated by an offline LLM and serve as fine-tuning labels. Meanwhile, we introduce MEMO, a novel metric that measures the model's memorization, to select optimal fine-tuning targets. The use of inverted facts not only maintains the covert nature of the model but also ensures that sensitive information is effectively forgotten without revealing the target data. Evaluated on the ToFU Knowledge Unlearning dataset using Llama2-7B-Chat and Phi-1.5, MEOW outperforms baselines in forgetting quality while preserving model utility. MEOW also maintains strong performance across NLU and NLG tasks and demonstrates superior resilience to attacks, validated via the Min-K% membership inference method.

PromptCoT: Synthesizing Olympiad-level Problems for Mathematical Reasoning in Large Language Models

Xueliang Zhao, Wei Wu, Jian Guan, Lingpeng Kong

The ability of large language models to solve complex mathematical problems has progressed significantly, particularly for tasks requiring advanced reasoning. However, the scarcity of sufficiently challenging problems, particularly at the Olympiad level, hinders further advancements. In this work, we introduce \textsc{PromptCoT}, a novel approach for automatically generating high-quality Olympiad-level math problems. The proposed method synthesizes complex problems based on mathematical concepts and the rationale behind problem construction, emulating the thought processes of experienced problem designers. We provide a theoretical analysis demonstrating that an optimal rationale should maximize both the likelihood of rationale generation given the associated concepts and the likelihood of problem generation conditioned on both the rationale and the concepts. Our method is evaluated on standard benchmarks including GSM8K, MATH-500, and AIME2024, where it consistently outperforms existing problem generation methods. Furthermore, we demonstrate that \textsc{PromptCoT} exhibits superior scalability, consistently maintaining high performance as the dataset size increases, outperforming the baselines.

Optimizing Reasoning for Text-to-SQL with Execution Feedback

Bohan Zhai, Canwen Xu, Zhewei Yao, Yuxiong He

Text-to-SQL demands precise reasoning to convert natural language questions into structured queries. While large language models (LLMs) excel in many reasoning tasks, their ability to leverage Chain-of-Thought (CoT) reasoning for text-to-SQL remains underexplored. We identify critical limitations: zero-shot CoT offers minimal gains, and Direct Preference Optimization (DPO) applied without CoT yields marginal improvements. We propose ExCoT-DPO, a novel framework that iteratively optimizes open-source LLMs by combining CoT reasoning with off-policy and on-policy DPO, relying solely on execution accuracy as feedback. This approach eliminates the need for reward models or human-annotated preferences. Our experimental results demonstrate significant performance gains: ExCoT-DPO improves execution accuracy on BIRD from 57.37% to 68.51% and on Spider from 78.81% to 86.59% for LLaMA-3 70B, with Qwen-2.5-Coder demonstrating similar improvements. Our best model achieves state-of-the-art performance in the single-model setting on both BIRD and Spider datasets.

Patterns Over Principles: The Fragility of Inductive Reasoning in LLMs under Noisy Observations

Chunyang Li, Weiqi Wang, Tianshi Zheng, Yangqiu Song

Inductive reasoning, a cornerstone of human cognition, enables generalization from limited data but hasn't yet been fully achieved by large language models (LLMs). While modern LLMs excel at reasoning tasks, their ability to maintain stable and consistent rule abstraction under imperfect observations remains underexplored. To fill this gap, in this work, we introduce Robust Rule Induction, a task that evaluates LLMs' capability in inferring rules from data that are fused with noisy examples. To address this task, we further propose Sample-steered Rule Refinement (SRR), a method enhancing reasoning stability via observation diversification and execution-guided feedback. Experiments across arithmetic, cryptography, and list functions reveal: (1) SRR outperforms other methods with minimal performance degradation under noise; (2) Despite slight accuracy variation, LLMs exhibit instability under noise (e.g., 0% accuracy change with only 70% consistent score); (3) Counterfactual task gaps highlight LLMs' reliance on memorized patterns over genuine abstraction. Our findings challenge LLMs' reasoning robustness, revealing susceptibility to hypothesis drift and pattern overfitting, while providing empirical evidence critical for developing human-like inductive systems. Code and data are available at https://github.com/HKUST-KnowComp/Robust-Rule-Induction.

EMGLLM: Data-to-Text Alignment for Electromyogram Diagnosis Generation with Medical Numerical Data Encoding

Zefei Long, Zhenbiao Cao, Wei Chen, zhongyu wei

Electromyography (EMG) tables are crucial for diagnosing muscle and nerve disorders, and advancing the automation of EMG diagnostics is significant for improving medical efficiency. EMG tables contain extensive continuous numerical data, which current Large Language Models (LLMs) often struggle to interpret effectively. To address this issue, we propose EMGLLM, a data-to-text model specifically designed for medical examination tables. EMGLLM employs the EMG Alignment Encoder to simulate the process that doctors compare test values with reference values, aligning the data into word embeddings that reflect health degree. Additionally, we construct ETM, a dataset comprising 17,276 real cases and their corresponding diagnostic results, to support medical data-to-text tasks. Experimental results on ETM demonstrate that EMGLLM outperforms various baseline models in understanding EMG tables and generating high-quality diagnoses, which represents an effective paradigm for automatic diagnosis generation from medical examination table.

Reverse Preference Optimization for Complex Instruction Following

Xiang Huang, Ting-En Lin, Feiteng Fang, Yuchuan Wu, Hangyu Li, Yuzhong Qu, Fei Huang, Yongbin Li

Instruction following (IF) is a critical capability for large language models (LLMs). However, handling complex instructions with multiple constraints remains challenging. Previous methods typically select preference pairs based on the number of constraints they satisfy, introducing noise where chosen examples may fail to follow some constraints and rejected examples may excel in certain respects over the chosen ones. To address the challenge of aligning with multiple preferences, we propose a simple yet effective method called Reverse Preference Optimization (RPO). It mitigates noise in preference pairs by dynamically reversing the constraints within the instruction to ensure the chosen response is perfect, alleviating the burden of extensive sampling and filtering to collect perfect responses. Besides, reversal also enlarges the gap between chosen and rejected responses, thereby clarifying the optimization direction and making it more robust to noise. We evaluate RPO on two multi-turn IF benchmarks, Sysbench and Multi-IF, demonstrating average improvements over the DPO baseline of 4.6 and 2.5 points(on Llama-3.1 8B), respectively. Moreover, RPO scales effectively across model sizes (8B to 70B parameters), with the 70B RPO model surpassing GPT-4o.

TechniqueRAG: Retrieval Augmented Generation for Adversarial Technique Annotation in Cyber Threat Intelligence Text

Ahmed Lekssays, Utsav Shukla, Husrev Taha Sencar, Md Rizwan Parvez

Accurately identifying adversarial techniques in security texts is critical for effective cyber defense. However, existing methods face a fundamental trade-off: they either rely on generic models with limited domain precision or require resource-intensive pipelines that depend on large labeled datasets and task-specific optimizations—such as custom hard-negative mining and denoising—resources rarely available in specialized domains. We propose TechniqueRAG, a domain-specific retrieval-augmented generation (RAG) framework that bridges this gap by integrating off-the-shelf retrievers, instruction-tuned LLMs, and minimal text–technique pairs. Our approach addresses data scarcity by fine-tuning only the generation component on limited in-domain examples, circumventing the need for resource-intensive retrieval training. While conventional RAG mitigates hallucination by coupling retrieval and generation, its reliance on generic retrievers often introduces noisy candidates, limiting domain-specific precision. To address this, we enhance retrieval quality and domain specificity through zero-shot LLM re-ranking, which explicitly aligns retrieved candidates with adversarial techniques. Experiments on multiple security benchmarks demonstrate that TechniqueRAG achieves state-of-the-art performance without extensive task-specific optimizations or labeled data, while comprehensive analysis provides further insights.

APT: Improving Specialist LLM Performance with Weakness Case Acquisition and Iterative Preference Training

Jun Rao, Zepeng Lin, Xuebo Liu, Xiaopeng Ke, Lian Lian, Dong Jin, shengjun cheng, Jun Yu, Min Zhang

Large Language Models (LLMs) often require domain-specific fine-tuning to address targeted tasks, which risks degrading their general capabilities. Maintaining a balance between domain-specific enhancements and general model utility is a key challenge. This paper proposes a novel approach named APT (Weakness Case Acquisition and Iterative Preference Training) to enhance domain-specific performance with self-generated dis-preferred weakness data (bad cases and similar cases). APT uniquely focuses on training the model using only those samples where errors occur, alongside a small, similar set of samples retrieved for this purpose. This targeted training minimizes interference with the model's existing knowledge base, effectively retaining generic capabilities. Experimental results on the LLama-2 and Mistral-V0.3 models across various benchmarks demonstrate that APT ensures no reduction in generic capacity and achieves superior performance on downstream tasks compared to various existing methods. This validates our method as an effective strategy for enhancing domain-specific capabilities without sacrificing the model's broader applicability.

Encode Errors: Representational Retrieval of In-Context Demonstrations for Multilingual Grammatical Error Correction

Guangyue Peng, Wei Li, Wen Luo, Houfeng Wang

Grammatical Error Correction (GEC) involves detecting and correcting incorrect usage of grammar. While large language models (LLMs) with in-context learning (ICL) capabilities have shown significant progress on various natural language processing (NLP) tasks, their few-shot performance on GEC remains suboptimal. This is mainly due to the challenge of retrieving suitable in-context demonstrations that capture error patterns instead of semantic similarity. In this paper, we demonstrate that LLMs can inherently capture information related to grammatical errors through their internal states. We extract from these states the Grammatical Error Representation (GER), an informative and semantically neutral encoding of grammatical errors. Our novel GER-based retrieval method significantly boosts performance in ICL settings on multilingual GEC datasets, improving the precision of correction. For high-resource languages, our results on 8B-sized open-source models match those of closed-source models such as Deepseek2.5 and GPT-4o-mini. For low-resource languages, our F_{0.5} scores surpass the baseline by a factor of 1.25. This method provides a more precise and resource-efficient solution for multilingual GEC, offering a promising direction for interpretable GEC research. The code will be available upon acceptance.

Measuring What Makes You Unique: Difference-Aware User Modeling for Enhancing LLM Personalization

Yilun Qiu, Xiaoyan Zhao, Yang Zhang, Yimeng Bai, Wenjie Wang, Hong Cheng, Fuli Feng, Tat-Seng Chua

Personalizing Large Language Models (LLMs) has become a critical step in facilitating their widespread application to enhance individual life experiences. In pursuit of personalization, distilling key preference information from an individual's historical data as instructional preference context to customize LLM generation has emerged as a promising direction. However, these methods face a fundamental limitation by overlooking the inter-user comparative analysis, which is essential for identifying the inter-user differences that truly shape preferences. To address this limitation, we propose Difference-aware Personalization Learning (DPL), a novel approach that emphasizes extracting inter-user differences to enhance LLM personalization. DPL strategically selects representative users for comparison and establishes a structured standard to extract meaningful, task-relevant differences for customizing LLM generation. Extensive experiments on real-world datasets demonstrate that DPL significantly enhances LLM personalization. We release our code on Anonymous GitHub https://anonymous.4open.science/r/DPL-2111.

A Classifier of Word-Level Variants in Witnesses of Biblical Hebrew Manuscripts

Iglika Nikolova-Stoupak, Maxime Amblard, Sophie Robert-Hayek, Davide D'Amico, Frédérique Rey

The current project is inscribed within the field of stemmatology or the study and/or reconstruction of textual transmission based on the relationship between the available witnesses of given texts. In particular, the variants (differences) at the word-level in manuscripts written in Biblical Hebrew are considered. A strong classifier (F1 value of 0.80) is trained to predict the category of difference between word pairs ('plus/minus', 'inversion', 'morphological', 'lexical' or 'unclassifiable') as present in collated (aligned) pairs of witnesses. The classifier is non-neural and makes use of the two words themselves as well as part-of-speech (POS) tags, hand-crafted rules per category and synthetically derived data. Other models experimented with include neural ones based on the state-of-the-art model for Modern Hebrew, DictaBERT. Other features whose relevance is tested are different types of morphological information pertaining to the word pairs and the Levenshtein distance between words. A selection of the strongest classifiers as well as the used synthetic data and the steps taken at its derivation are made available. Coincidentally, the corelation between two sets of morphological labels is investigated: professionally established as per the Qumran-Digital online library and automatically derived with the sub-model DictaBERT-morph.

NOVA: An Iterative Planning Framework for Enhancing Scientific Innovation with Large Language Models

xiang hu, Hongyu Fu, Jinge Wang, Yifeng wang, zhikun li, Renjun Xu, Yu Lu, Yaochu Jin, Lili Pan, Zhenzhong Lan

Scientific innovation is pivotal for humanity, and harnessing large language models (LLMs) to generate research ideas could transform discovery. However, existing LLMs often produce simplistic and repetitive suggestions due to their limited ability in acquiring external knowledge for innovation. To address this problem, we introduce an enhanced planning and search methodology designed to boost the creative potential of LLM-based systems. Our approach involves an iterative process to purposely plan the retrieval of external knowledge, progressively enriching the idea generation with broader and deeper insights. Validation through automated and human assessments demonstrates that our framework substantially elevates the quality of generated ideas, particularly in novelty and diversity. The number of unique novel ideas produced by our framework is 3.4 times higher than without it. Moreover, our method outperforms the current state-of-the-art, generating at least 2.5 times more top-rated ideas based on 170 seed papers in a Swiss Tournament evaluation. Our code is available in [Anonymous Github Repository](https://anonymous.4open.science/r/Nova-4CA8/README.md) Repository. A demo of our work can be found at [Nova](https://www.nova.vin/).

PEToolLLM: Towards Personalized Tool Learning in Large Language Models

Qiancheng Xu, Yongqi Li, Heming Xia, Fan Liu, Min Yang, Wenjie Li

Tool learning has emerged as a promising direction by extending Large Language Models' (LLMs) capabilities with external tools. Existing tool learning studies primarily focus on the general-purpose tool-use capability, which addresses explicit user requirements in instructions. However, they overlook the importance of personalized tool-use capability, leading to an inability to handle implicit user preferences. To address the limitation, we first formulate the task of personalized tool learning, which integrates user's interaction history towards personalized tool usage. To fill the gap of missing benchmarks, we construct PEToolBench, featuring diverse user preferences reflected in interaction history under three distinct personalized settings, and encompassing a wide range of tool-use scenarios. Moreover, we propose a framework PEToolLLaMA to adapt LLMs to the personalized tool learning task, which is trained through supervised fine-tuning and direct preference optimization. Extensive experiments on PEToolBench demonstrate the superiority of PEToolLLaMA over existing LLMs. We release code and data for review at https://anonymous.4open.science/r/PEToolBench-952F/.

gMBA: Expression Semantic Guided Mixed Boolean-Arithmetic Deobfuscation Using Transformer Architectures

Youjeong Roh, Joon-Young Paik, Jingun Kwon, Eun-Sun Cho

Mixed Boolean-Arithmetic (MBA) obfuscation protects intellectual property by converting programs into forms that are more complex to analyze. However, MBA has been increasingly exploited by malware developers to evade detection and cause significant real-world problems. Traditional MBA deobfuscation methods often consider these expressions as part of a black box and overlook their internal semantic information. To bridge this gap, we propose a truth table, which is an automatically constructed semantic representation of an expression's behavior that does not rely on external resources. The truth table is a mathematical form that represents the output of expression for all possible combinations of input. We also propose a general and extensible guided MBA deobfuscation framework (gMBA) that modifies a Transformer-based neural encoder-decoder Seq2Seq architecture to incorporate this semantic guidance. Experimental results and in-depth analysis show that integrating expression semantics significantly improves performance and highlights the importance of internal semantic expressions in recovering obfuscated code to its original form.

TicTac: Temporal-aware Supervised Fine-tuning for Automatic Text Dating

Minna Peng, Han Ren

Pre-trained language models have achieved success in many natural language processing tasks, whereas they are trapped by the time-agnostic setting, impacting the performance in automatic text dating. Studies on adopting sophisticated learning models to capture the lexical knowledge of semantic change which helps improve task performance, whereas few of them take into account the temporal relatedness of documents, which may be helpful for learning temporal semantic variations. This paper introduces TicTac, a supervised fine-tuning model for text dating. Based on a fine-tuning framework, the model employs a contrastive learning-based document modeling approach, trying to capture temporal implications in time-ordered documents. The model also adopts a metric learning approach to improve model performances by appropriately correcting prediction errors of different degrees. Experiments on two diachronic corpora show that our model effectively captures the temporal semantic information and outperforms state-of-the-art baselines.

Label-semantics Aware Generative Approach for Domain-Agnostic Multilabel Classification

Subhendu Khatuya, Shashwat Naidu, Saptarshi Ghosh, Pawan Goyal, Niloy Ganguly

The explosion of textual data has made manual document classification increasingly challenging. To address this, we introduce a robust, efficient domain-agnostic generative model framework for multi-label text classification. Instead of treating labels as mere atomic symbols, our approach utilizes predefined label descriptions and is trained to generate these descriptions based on the input text. During inference, the generated descriptions are matched to the predefined labels using a finetuned sentence transformer. We integrate this with a dual-objective loss function, combining cross-entropy loss and cosine similarity of the generated sentences with the predefined target descriptions, ensuring both semantic alignment and accuracy. Our proposed model LAGAMC stands out for its parameter efficiency and versatility across diverse datasets, making it well-suited for practical applications. We demonstrate the effectiveness of our proposed model by achieving new state-of-the-art performances across all evaluated datasets, surpassing several strong baselines. We achieve improvements of 13.94 % in Micro-F1 and 24.85 % in Macro-F1 compared to the closest baseline across all datasets.

M2PA: A Multi-Memory Planning Agent for Open Worlds Inspired by Cognitive Theory

YanfangZhou, Xiaodong Li, Yuntao Liu, Yongqiang Zhao, Xintong Wang, Zhenyu Li, Jinlong Tian, Xinhai Xu

Open-world planning poses a significant challenge for general artificial intelligence due to environmental complexity and task diversity, especially in long-term tasks and lifelong learning. Inspired by cognitive theories, we propose M2PA, an open-world multi-memory planning agent. M2PA innovates by combining Large Language Models (LLMs) with human-like multi-memory systems, aiming to fully leverage the strengths of both while mitigating their respective limitations. By integrating the expansive world knowledge and language processing capabilities of LLMs with the perception and experience accumulation abilities of the human memory system, M2PA exhibits situation awareness, and experience generalization capabilities, as well as the potential for lifelong learning. In experiments, M2PA significantly outperforms current state-of-the-art agents across 50 Minecraft tasks in zero-shot learning. In exploratory lifelong learning experiments, M2PA demonstrates its continuous learning ability, achieving a \textbf{38.33%} success rate in the “ObtainDiamond” task. Our findings provide a novel paradigm for constructing more effective agents in open-world environments.

Thinking Before Running! Efficient Code Generation with Thorough Exploration and Optimal Refinement

Xiaoqing Zhang, Yuhan Liu, Flood Sung, Xiuying Chen, Shuo Shang, Rui Yan

Code generation is crucial in software engineering for automating the coding process efficiently. While test-time computation methods show promise, they suffer from high latency due to multiple computation rounds. To overcome this, we introduce \textbf{ThinkCoder}, a framework that combines thorough exploration with optimal refinement. The exploration phase diversifies the solution space by searching for potential solutions, followed by a refinement phase that enhances precision. This approach allows us to select the best solution through careful consideration before taking action, avoiding excessive trial and error. To further minimize test-time computation overhead, we introduce preference-driven optimization with Reinforced Self-Training (ReST), which uses exploration trajectories from ThinkCoder to guide LLM's evolution. This approach enhances LLM's exploration efficiency via preference learning, cutting costs while maintaining accuracy. ThinkCoder boosts the performance with a single LLM, excelling on benchmarks like HumanEval and MBPP. Compared to SOTA models, it improves Pass@1 by 3.0\% over MapCoder with just 6.4\% of the computation cost. Against AgentCoder, ThinkCoder achieves a 0.5\% higher Pass@1 after 2 rounds, outperforming AgentCoder's 5 rounds. Additionally, ReST with success trajectories enhances efficiency, allowing models like LLaMA2-7B to achieve competitive results using only 20\% of the computational resources. These results highlight the framework's effectiveness and scalability. \footnote{The code is available at }

PodAgent: A Comprehensive Framework for Podcast Generation

Yujia Xiao, Lei He, Haohan Guo, Feng-Long Xie, Tan Lee

Existing automatic audio generation methods struggle to generate podcast-like audio programs effectively. The key challenges lie in in-depth content generation, appropriate and expressive voice production. This paper proposed PodAgent, a comprehensive framework for creating audio programs. PodAgent 1) generates informative topic-discussion content by designing a Host-Guest-Writer multi-agent collaboration system, 2) builds a voice pool for suitable voice-role matching and 3) utilizes LLM-enhanced speech synthesis method to generate expressive conversational speech. Given the absence of standardized evaluation criteria for podcast-like audio generation, we developed comprehensive assessment guidelines to effectively evaluate the model's performance. Experimental results demonstrate PodAgent's effectiveness, significantly surpassing direct GPT-4 generation in topic-discussion dialogue content, achieving an 87.4\% voice-matching accuracy, and producing more expressive speech through LLM-guided synthesis. Demo page: https://podcast-agent.github.io/demo/. Source code: https://github.com/yujxx/PodAgent.

STORM-BORN: A Challenging Mathematical Derivations Dataset Curated via a Human-in-the-Loop Multi-Agent Framework

Wenhao Liu, Zhenyi Lu, Xinyu Hu, Jerry Zhang, Dailin Li, Jiacheng Cen, Huilin Cao, Haiteng Wang, Yuhan Li, XIE KUN, Dandan Li, Pei Zhang, Chengbo Zhang, Yuxiang Ren, Xiaohong Huang, Yan Ma

High-quality math datasets are crucial for advancing the reasoning abilities of large language models (LLMs). However, existing datasets often suffer from three key issues: outdated and insufficient challenging content, neglecting human-like reasoning, and limited reliability due to single-LLM generation. To address these, we introduce \textbf{STORM-BORN}, an ultra-challenging dataset of mathematical derivations sourced from cutting-edge academic papers, which includes dense human-like approximations and heuristic cues. To ensure the reliability and quality, we propose a novel human-in-the-loop, multi-agent data generation framework, integrating reasoning-dense filters, multi-agent collaboration, and human mathematicians' evaluations. We curated a set of 2,000 synthetic samples and deliberately selected the 100 most difficult problems. Even most advanced models like GPT-o1 solved fewer than 5\% of them. Fine-tuning on STORM-BORN boosts accuracy by 7.84\% (LLaMA3-8B) and 9.12\% (Qwen2.5-7B). As AI approaches mathematician-level reasoning, STORM-BORN provides both a high-difficulty benchmark and a human-like reasoning training resource. Our code and dataset are publicly available at https://github.com/lwhere/STORM-BORN.

Memory-augmented Query Reconstruction for LLM-based Knowledge Graph Reasoning

Mufan Xu, Gewen Liang, Kehai Chen, wei wang, Xun Zhou, Muyun Yang, Tiejun Zhao, Min Zhang

Large language models (LLMs) have achieved remarkable performance on knowledge graph question answering (KGQA) tasks by planning and interacting with knowledge graphs. However, existing methods often confuse tool utilization with knowledge reasoning, harming readability of model outputs and giving rise to hallucinatory tool invocations, which hinder the advancement of KGQA. To address this issue, we propose Memory-augmented Query Reconstruction for LLM-based Knowledge Graph Reasoning (MemQ) to decouple LLM from tool invocation tasks using LLM-built query memory. By establishing a memory module with explicit descriptions of query statements, the proposed MemQ facilitates the KGQA process with natural language reasoning and memory-augmented query reconstruction. Meanwhile, we design an effective and readable reasoning to enhance the LLM's reasoning capability in KGQA. Experimental results that MemQ achieves state-of-the-art performance on widely used benchmarks WebQSP and CWQ.

MOSAIC: Multiple Observers Spotting AI Content

Matthieu Dubois, François Yvon, Pablo Piantanida

The dissemination of Large Language Models (LLMs), trained at scale, and endowed with powerful text-generating abilities, has made it easier for all to produce harmful, toxic, faked or forged content. In response, various proposals have been made to automatically discriminate artificially generated from human-written texts, typically framing the problem as a binary classification problem. Early approaches evaluate an input document with a well-chosen detector LLM, assuming that low-perplexity scores reliably signal machine-made content. More recent systems instead consider two LLMs and compare their probability distributions over the document to further discriminate when perplexity alone cannot. However, using a fixed pair of models can induce brittleness in performance. We extend these approaches to the ensembling of several LLMs and derive a new, theoretically grounded approach to combine their respective strengths. Our experiments, using a variety of generator LLMs, suggest that this approach effectively harnesses each model’s capabilities, leading to strong detection performance on a variety of domains.

Automated main concept generation for narrative discourse assessment in aphasia

Ankita Gupta, Marisa Hudspeth, Jacquie Kurland, Brendan O'Connor

We present an interesting application of narrative understanding in the clinical assessment of aphasia, where story retelling tasks are used to evaluate a patient's communication abilities. This clinical setting provides a framework to help operationalize narrative discourse analysis and an application-focused evaluation method for narrative understanding systems. In particular, we highlight the use of main concepts---a list of statements that capture a story's gist---for aphasic discourse analysis and propose the task of automatically generating MCs from novel stories, that can be edited manually by experts thus enabling wider adaptation of current assessment tools. We then develop a prompt ensemble method using large language models (LLMs) to automatically generate MCs for a given story. We evaluate our proposed method on an existing narrative summarization dataset and apply it to a set of stories that have been annotated with main concepts through extensive analysis of retells from non-aphasic and aphasic participants (Kurland et al., 2021). Our results show that our method can generate most of the gold-standard MCs for stories from this dataset. Finally, we release this dataset of these stories and their annotated MCs, to spur more research in this area.

Enhancing LLM-based Hatred and Toxicity Detection with Meta-Toxic Knowledge Graph

Yibo Zhao, Jiapeng Zhu, Can Xu, Yao Liu, Xiang Li

The rapid growth of social media platforms has raised significant concerns regarding online content toxicity. When Large Language Models (LLMs) are used for toxicity detection, two key challenges emerge: 1) the absence of domain-specific toxicity knowledge leads to false negatives; 2) the excessive sensitivity of LLMs to toxic speech results in false positives, limiting freedom of speech. To address these issues, we propose a novel method called *MetaTox*, leveraging graph search on a meta-toxic knowledge graph to enhance hatred and toxicity detection. First, we construct a comprehensive meta-toxic knowledge graph by utilizing LLMs to extract toxic information through a three step pipeline. Second, we query the graph via retrieval and ranking processes to supplement accurate, relevant toxicity knowledge. Extensive experiments and case studies across multiple datasets demonstrate that our MetaTox boosts overall toxicity detection performance, particularly in out-of-domain settings. In addition, under in-domain scenarios, we surprisingly find that small language models are more competent. Our code is available at https://github.com/YiboZhao624/MetaTox.

Answer When Needed, Forget When Not: Language Models Pretend to Forget via In-Context Knowledge Unlearning

Shota Takashiro, Takeshi Kojima, Andrew Gambardella, Qi Cao, Yusuke Iwasawa, Yutaka Matsuo

As large language models (LLMs) are applied across diverse domains, the ability to selectively unlearn specific information has become increasingly essential. For instance, LLMs are expected to provide certain confidential information to authorized internal users, such as employees or trusted partners, while withholding it from external users, including the general public and unauthorized entities. In response to this challenge, we propose a novel method termed "in-context knowledge unlearning", which enables the model to selectively forget information in test-time based on the query context. Our method fine-tunes pre-trained LLMs to enable prompt unlearning of target knowledge within the context, while preserving other knowledge. Experiments on TOFU, AGE and RWKU datasets using Llama2-7B/13B and Mistral-7B models show that our method achieves up to 95% forget accuracy while retaining 80% of unrelated knowledge, significantly outperforming baselines in both in-domain and out-of-domain scenarios. Further investigation of the model's internal behavior revealed that while fine-tuned LLMs generate correct predictions in the middle layers and maintain them up to the final layer, they make the decision to forget at the last layer, i.e. "LLMs pretend to forget". Our findings offer valuable insight into the improvement of the robustness of the unlearning mechanisms in LLM, setting a foundation for future research in the field.

LLMs as Planning Modelers: A Survey for Leveraging Large Language Models to Construct Automated Planning Models

Marcus Tantakoun, Christian Muise, Xiaodan Zhu

Large Language Models (LLMs) excel in various natural language tasks but often struggle with long-horizon planning problems requiring structured reasoning. This limitation has drawn interest in integrating neuro-symbolic approaches within the Automated Planning (AP) and Natural Language Processing (NLP) communities. However, identifying optimal AP deployment frameworks can be daunting. This paper aims to provide a timely survey of the current research with an in-depth analysis, positioning LLMs as tools for extracting and refining planning models to support reliable AP planners. By systematically reviewing the current state of research, we highlight methodologies, and identify critical challenges and future directions, hoping to contribute to the joint research on NLP and Automated Planning.

IMPARA-GED: Grammatical Error Detection is Boosting Reference-free Grammatical Error Quality Estimator

Yusuke Sakai, Takumi Goto, Taro Watanabe

We propose IMPARA-GED, a novel reference-free automatic grammatical error correction (GEC) evaluation method with grammatical error detection (GED) capabilities. We focus on the quality estimator of IMPARA, an existing automatic GEC evaluation method, and construct that of IMPARA-GED using a pre-trained language model with enhanced GED capabilities. Experimental results on SEEDA, a meta-evaluation dataset for automatic GEC evaluation methods, demonstrate that IMPARA-GED achieves the highest correlation with human sentence-level evaluations.

Low-Resource Grammatical Error Correction: Selective Data Augmentation with Round-Trip Machine Translation

Frank Palma Gomez, Alla Rozovskaya

State-of-the-art methods for grammatical error correction require large amounts of parallel data for training. Due to lack of gold-labeled data, techniques that create synthetic training data have become popular. We show that models trained on such data tend to correct a limited range of grammar and spelling mistakes that involve character-level changes, but perform poorly on (more complex) phenomena that require word-level changes. We propose to address the gap in performance on such errors by generating synthetic data through selective data augmentation via round-trip machine translation. We show that the proposed technique, SeLex-RT is capable of generating mistakes that are similar to those observed with language learners. Using the approach with two types of state-of-the-art learning frameworks and two languages (Russian and Ukrainian), we achieve substantial improvements, compared to training on synthetic data produced with standard techniques. The method is especially effective in low-resource settings. Analysis of the output reveals that models trained on data noisified using round-trip translation are capable of making word-level changes and correct lexical errors common with language learners.

Knowledge Base Construction for Knowledge-Augmented Text-to-SQL

Jinheon Baek, Horst Samulowitz, Oktie Hassanzadeh, Dharmashankar Subramanian, Sola Shirai, Alfio Gliozzo, Debarun Bhattacharjya

Text-to-SQL aims to translate natural language queries into SQL statements, which is practical as it enables anyone to easily retrieve the desired information from databases. Recently, many existing approaches tackle this problem with Large Language Models (LLMs), leveraging their strong capability in understanding user queries and generating corresponding SQL code. Yet, the parametric knowledge in LLMs might be limited to covering all the diverse and domain-specific queries that require grounding in various database schemas, which makes generated SQLs less accurate oftentimes. To tackle this, we propose constructing the knowledge base for text-to-SQL, a foundational source of knowledge, from which we retrieve and generate the necessary knowledge for given queries. In particular, unlike existing approaches that either manually annotate knowledge or generate only a few pieces of knowledge for each query, our knowledge base is comprehensive, which is constructed based on a combination of all the available questions and their associated database schemas along with their relevant knowledge, and can be reused for unseen databases from different datasets and domains. We validate our approach on multiple text-to-SQL datasets, considering both the overlapping and non-overlapping database scenarios, where it outperforms relevant baselines substantially.

From Complexity to Clarity: AI/NLP's Role in Regulatory Compliance

Jivitesh Jain, Nivedhitha Dhanasekaran, Mona T. Diab

Regulatory compliance is a cornerstone of trust and accountability in critical sectors like finance, healthcare, and technology, yet its complexity poses significant challenges for organizations worldwide. Recent advances in natural language processing, particularly large language models, have demonstrated remarkable capabilities in text analysis and reasoning, offering promising solutions for automating compliance processes. This survey examines the current state of automated compliance, analyzing key challenges and approaches across problem areas. We identify critical limitations in current automated compliance datasets and techniques, including issues of adaptability, completeness, and trust. Looking ahead, we propose research directions to address these challenges, emphasizing standardized evaluation frameworks and balanced human-AI collaboration.

How Programming Concepts and Neurons Are Shared in Code Language Models

Amir Hossein Kargaran, Yihong Liu, François Yvon, Hinrich Schuetze

Several studies have explored the mechanisms of large language models (LLMs) in coding tasks, but most have focused on programming languages (PLs) in a monolingual setting. In this paper, we investigate the relationship between multiple PLs and English in the concept space of LLMs. We perform a few-shot translation task on 21 PL pairs using two Llama-based models. By decoding the embeddings of intermediate layers during this task, we observe that the concept space is closer to English (including PL keywords) and assigns higher probabilities to English tokens in the second half of the middle layers, but not in the top layers. We analyze neuron activations for 11 PLs and English, finding that while language-specific neurons are primarily concentrated in the bottom layers, those exclusive to each PL tend to appear in the top layers. For PLs that are highly aligned with multiple other PLs, identifying language-specific neurons is not feasible. These PLs also tend to have a larger keyword set than other PLs and are closer to the model's concept space regardless of the input/output PL in translation task. Our findings provide insights into how LLMs internally represent PLs, revealing structural patterns in the model's concept space. Code is available at [anon. URL].

Automating Legal Concept Interpretation with LLMs: Retrieval, Generation, and Evaluation

Kangcheng Luo, Quzhe Huang, Cong Jiang, Yansong Feng

Interpreting the law is always essential for the law to adapt to the ever-changing society. It is a critical and challenging task even for legal practitioners, as it requires meticulous and professional annotations and summarizations by legal experts, which are admittedly time-consuming and expensive to collect at scale. To alleviate the burden on legal experts, we propose a method for automated legal interpretation. Specifically, by emulating doctrinal legal research, we introduce a novel framework, **ATRIE**, to address Legal Concept Interpretation, a typical task in legal interpretation. **ATRIE** utilizes large language models (LLMs) to **A**u**T**omatically **R**etrieve concept-related information, **I**nterpret legal concepts, and **E**valuate generated interpretations, eliminating dependence on legal experts. ATRIE comprises a legal concept interpreter and a legal concept interpretation evaluator. The interpreter uses LLMs to retrieve relevant information from previous cases and interpret legal concepts. The evaluator uses performance changes on Legal Concept Entailment, a downstream task we propose, as a proxy of interpretation quality. Automated and multifaceted human evaluations indicate that the quality of our interpretations is comparable to those written by legal experts, with superior comprehensiveness and readability. Although there remains a slight gap in accuracy, it can already assist legal practitioners in improving the efficiency of legal interpretation.

DeepSolution: Boosting Complex Engineering Solution Design via Tree-based Exploration and Bi-point Thinking

Zhuoqun Li, Haiyang Yu, Xuanang Chen, Hongyu Lin, Yaojie Lu, Fei Huang, Xianpei Han, Yongbin Li, Le Sun

Designing solutions for complex engineering challenges is crucial in human production activities. However, previous research in the retrieval-augmented generation (RAG) field has not sufficiently addressed tasks related to the design of complex engineering solutions. To fill this gap, we introduce a new benchmark, SolutionBench, to evaluate a system's ability to generate complete and feasible solutions for engineering problems with multiple complex constraints. To further advance the design of complex engineering solutions, we propose a novel system, SolutionRAG, that leverages the tree-based exploration and bi-point thinking mechanism to generate reliable solutions. Extensive experimental results demonstrate that SolutionRAG achieves state-of-the-art (SOTA) performance on the SolutionBench, highlighting its potential to enhance the automation and reliability of complex engineering solution design in real-world applications.

Explicit and Implicit Data Augmentation for Social Event Detection

Congbo Ma, Yuxia Wang, Jia Wu, Jian Yang, Jing Du, Zitai Qiu, Qing Li, Hu Wang, Preslav Nakov

Social event detection involves identifying and categorizing important events from social media, which relies on labeled data, but annotation is costly and labor-intensive. To address this problem, we propose Augmentation framework for Social Event Detection (SED-Aug), a plug-and-play dual augmentation framework, which combines explicit text-based and implicit feature-space augmentation to enhance data diversity and model robustness. The explicit augmentation utilizes LLMs to enhance textual information through five diverse generation strategies. For implicit augmentation, we design five novel perturbation techniques that operate in the feature space on structural fused embeddings. These perturbations are crafted to keep the semantic and relational properties of the embeddings and make them more diverse. Specifically, SED-Aug outperforms the best baseline model by approximately 17.67% on the Twitter2012 dataset and by about 15.57% on the Twitter2018 dataset in terms of the average F1 score.

AIMSCheck: Leveraging LLMs for AI-Assisted Review of Modern Slavery Statements Across Jurisdictions

Adriana Eufrosina Bora, Akshatha Arodi, Duoyi Zhang, Jordan Bannister, Mirko Bronzi, Arsene Fansi Tchango, Md Abul Bashar, Richi Nayak, Kerrie Mengersen

Modern Slavery Acts mandate that corporations disclose their efforts to combat modern slavery, aiming to enhance transparency and strengthen practices for its eradication. However, verifying these statements remains challenging due to their complex, diversified language and the sheer number of statements that must be reviewed. The development of NLP tools to assist in this task is also difficult due to a scarcity of annotated data. Furthermore, as modern slavery transparency legislation has been introduced in several countries, the generalizability of such tools across legal jurisdictions must be studied. To address these challenges, we work with domain experts to make two key contributions. First, we present AIMS.uk and AIMS.ca, newly annotated datasets from the UK and Canada to enable cross-jurisdictional evaluation. Second, we introduce AIMSCheck, an end-to-end framework for compliance validation. AIMSCheck decomposes the compliance assessment task into three levels, enhancing interpretability and practical applicability. Our experiments show that models trained on an Australian dataset generalize well across UK and Canadian jurisdictions, demonstrating the potential for broader application in compliance monitoring. We release the benchmark datasets and AIMSCheck to the public to advance AI-adoption in compliance assessment and drive further research in this field.

TeRDy: Temporal Relation Dynamics through Frequency Decomposition for Temporal Knowledge Graph Completion

Ziyang Liu, Chaokun Wang

Temporal knowledge graph completion aims to predict missing facts in a knowledge graph by leveraging temporal information. Existing methods often struggle to capture both the long-term changes and short-term variability of relations, which are crucial for accurate prediction. In this paper, we propose a novel method called TeRDy for temporal knowledge graph completion. TeRDy captures temporal relational dynamics by utilizing time-invariant embeddings, along with long-term temporally dynamic embeddings (e.g., enduring political alliances) and short-term temporally dynamic embeddings (e.g., transient political events). These two types of embeddings are derived from low- and high-frequency components via frequency decomposition. Also, we design temporal smoothing and temporal gradient to seamlessly incorporate timestamp embeddings into relation embeddings. Extensive experiments on benchmark datasets demonstrate that TeRDy outperforms state-of-the-art temporal knowledge graph embedding methods.

SHARE: An SLM-based Hierarchical Action CorREction Assistant for Text-to-SQL

Ge Qu, Jinyang Li, Bowen Qin, Xiaolong Li, Nan Huo, Chenhao Ma, Reynold Cheng

Current self-correction approaches in text-to-SQL face two critical limitations: 1) Conventional self-correction methods rely on recursive self-calls of LLMs, resulting in multiplicative computational overhead, and 2) LLMs struggle to implement effective error detection and correction for monolithic SQL queries, as they fail to demonstrate the underlying reasoning path. In this work, we propose **SHARE**, a **S**LM-based **H**ierarchical **A**ction cor**RE**ction assistant that enables LLMs to perform more precise error localization and efficient correction. SHARE orchestrates three specialized Small Language Models (SLMs) in a sequential pipeline, where it first transforms monolithic SQL queries into stepwise action trajectories that reveal underlying reasoning, followed by a two-phase granular refinement. We further propose a novel hierarchical self-evolution strategy for data-efficient training. Our experimental results demonstrate that SHARE effectively enhances self-correction capabilities while proving robust across various LLMs. Furthermore, our comprehensive analysis shows that SHARE maintains strong performance even in low-resource training settings, which is particularly valuable for text-to-SQL applications with data privacy constraints.

HomeBench: Evaluating LLMs in Smart Homes with Valid and Invalid Instructions Across Single and Multiple Devices

Silin Li, Yuhang Guo, Jiashu Yao, Zeming Liu, Haifeng Wang

Large language models (LLMs) have the potential to revolutionize smart home assistants by enhancing their ability to accurately understand user needs and respond appropriately, which is extremely beneficial for building a smarter home environment. While recent studies have explored integrating LLMs into smart home systems, they primarily focus on handling straightforward, valid single-device operation instructions. However, real-world scenarios are far more complex and often involve users issuing invalid instructions or controlling multiple devices simultaneously. These have two main challenges: LLMs must accurately identify and rectify errors in user instructions and execute multiple user instructions perfectly. To address these challenges and advance the development of LLM-based smart home assistants, we introduce HomeBench, the first smart home dataset with valid and invalid instructions across single and multiple devices in this paper. We have experimental results on 13 distinct LLMs; e.g., GPT-4o achieves only a 0.0\% success rate in the scenario of invalid multi-device instructions, revealing that the existing state-of-the-art LLMs still cannot perform well in this situation even with the help of in-context learning, retrieval-augmented generation, and fine-tuning. Our code and dataset are publicly available at https://github.com/BITHLP/HomeBench.

Planning-Driven Programming: A Large Language Model Programming Workflow

Chao Lei, Yanchuan Chang, Nir Lipovetzky, Krista A. Ehinger

The strong performance of large language models (LLMs) raises extensive discussion on their application to code generation. Recent research suggests continuous program refinements through visible tests to improve code generation accuracy in LLMs. However, these methods suffer from LLMs' inefficiency and limited reasoning capacity. In this work, we propose an LLM programming workflow (LPW) designed to improve both initial code generation and subsequent refinements within a structured two-phase workflow. Specifically, the solution generation phase formulates a solution plan, which is then verified through visible tests to specify the intended natural language solution. Subsequently, the code implementation phase drafts an initial code according to the solution plan and its verification. If the generated code fails the visible tests, the plan verification serves as the intended solution to consistently inform the refinement process for correcting bugs. Compared to state-of-the-art methods across various existing LLMs, LPW significantly improves the Pass@1 accuracy by up to 16.4% on well-established text-to-code generation benchmarks. LPW also sets new state-of-the-art Pass@1 accuracy, achieving 98.2% on HumanEval, 84.8% on MBPP, 59.3% on LiveCode, 62.6% on APPS, and 34.7% on CodeContests, using GPT-4o as the backbone. Our code is publicly available at: https://github.com/you68681/lpw.

Knowledge Decoupling via Orthogonal Projection for Lifelong Editing of Large Language Models

Haoyu Xu, Pengxiang Lan, Enneng Yang, Guibing Guo, Jianzhe Zhao, Linying Jiang, Xingwei Wang

As large language models (LLMs) require continuous knowledge updates and the mitigation of hallucination issues in generated content, lifelong model editing has become a prominent research area. A mainstream knowledge editing method usually freezes LLM's original parameters and adds extra trainable modules for new knowledge management, reducing interference with old knowledge. Although these approaches have achieved some success, our experiments show that, after extensive editing, the model's knowledge understanding and memory capacity significantly degrade, particularly concerning early edited knowledge. The root cause is that subsequent edits interfere with the previously edited knowledge, and we refer to this phenomenon as knowledge coupling. To address this issue, we propose the \textbf{\underline{K}nowledge \underline{D}ecoupling \underline{E}diting} (KDE) method. Specifically, KDE stores the basis vectors of the representation space of past edits in a knowledge cache. It projects the gradient of the current edit onto a space orthogonal to previous knowledge for updating. This method effectively alleviates the coupling between different pieces of knowledge. We also propose a two-stage training strategy to better balance the model's ability to edit new knowledge and distinguish whether a query is related to previous edits. This strategy gradually reduces the interference between new knowledge editing and query distinction, maintaining stable performance during long-term editing. We compared KDE with nine cutting-edge editing methods across multiple mainstream LLMs. The results demonstrate that, regarding question-answering ability and hallucination mitigation, KDE achieves average improvements of 14% and 61%.

-Decoding: Adaptive Foresight Sampling for Balanced Inference-Time Exploration and Exploitation

Fangzhi Xu, Hang Yan, Chang Ma, Haiteng Zhao, Jun Liu, Qika Lin, Zhiyong Wu

Inference-time optimization scales computation to derive deliberate reasoning steps for effective performance. While previous search-based strategies address the short-sightedness of auto-regressive generation, the vast search space leads to excessive exploration and insufficient exploitation. To strike an efficient balance to derive the optimal step, we frame the decoding strategy as foresight sampling, leveraging simulated future steps to obtain globally optimal step estimation. Built on it, we propose a novel decoding strategy, named \phi-Decoding. To provide a precise and expressive estimation of step value, \phi-Decoding approximates two distributions via foresight and clustering. Sampling from the joint distribution, the optimal steps can be selected for exploitation. To support adaptive computation allocation, we propose in-width and in-depth pruning strategies, featuring a light-weight solution to achieve inference efficiency. Extensive experiments across seven benchmarks show \phi-Decoding outperforms strong baselines in both performance and efficiency. Additional analysis demonstrates its generalization across various LLMs and scalability across a wide range of computing budgets.

MEDDxAgent: A Unified Modular Agent Framework for Explainable Automatic Differential Diagnosis

Daniel Philip Rose, Chia-Chien Hung, Marco Lepri, Israa Alqassem, Kiril Gashteovski, Carolin Lawrence

Differential Diagnosis (DDx) is a fundamental yet complex aspect of clinical decision-making, in which physicians iteratively refine a ranked list of possible diseases based on symptoms, antecedents, and medical knowledge. While recent advances in large language models (LLMs) have shown promise in supporting DDx, existing approaches face key limitations, including single-dataset evaluations, isolated optimization of components, unrealistic assumptions about complete patient profiles, and single-attempt diagnosis. We introduce a Modular Explainable DDx Agent (MEDDxAgent) framework designed for interactive DDx, where diagnostic reasoning evolves through iterative learning, rather than assuming a complete patient profile is accessible. MEDDxAgent integrates three modular components: (1) an orchestrator (DDxDriver), (2) a history taking simulator, and (3) two specialized agents for knowledge retrieval and diagnosis strategy. To ensure robust evaluation, we introduce a comprehensive DDx benchmark covering respiratory, skin, and rare diseases. We analyze single-turn diagnostic approaches and demonstrate the importance of iterative refinement when patient profiles are not available at the outset. Our broad evaluation demonstrates that MEDDxAgent achieves over 10% accuracy improvements in interactive DDx across both large and small LLMs, while offering critical explainability into its diagnostic reasoning process.

A Training-free LLM-based Approach to General Chinese Character Error Correction

Houquan Zhou, Bo Zhang, Zhenghua Li, Ming Yan, Min Zhang

Chinese spelling correction (CSC) is a crucial task that aims to correct character errors in Chinese text. While conventional CSC focuses on character substitution errors caused by mistyping, two other common types of character errors, missing and redundant characters, have received less attention. These errors are often excluded from CSC datasets during the annotation process or ignored during evaluation, even when they have been annotated. This issue limits the practicality of the CSC task. To address this issue, we introduce the task of General Chinese Character Error Correction (C2EC), which focuses on all three types of character errors. We construct a high-quality C2EC benchmark by combining and manually verifying data from CCTC and Lemon datasets. We extend the training-free prompt-free CSC method to C2EC by using Levenshtein distance for handling length changes and leveraging an additional prompt-based large language model (LLM) to improve performance. Experiments show that our method enables a 14B-parameter LLM to be on par with models nearly 50 times larger on both conventional CSC and C2EC tasks, without any fine-tuning.

One for All: Update Parameterized Knowledge Across Multiple Models with Once Edit

Weitao Ma, Xiyuan Du, Xiaocheng Feng, Lei Huang, Yichong Huang, Huiyi Zhang, Xiaoliang Yang, Baohang Li, Xiachong Feng, Ting Liu, Bing Qin

Large language models (LLMs) encode vast world knowledge but struggle to stay up-to-date, often leading to errors and hallucinations. Knowledge editing offers an efficient alternative to retraining, enabling targeted modifications by updating specific model parameters. However, existing methods primarily focus on individual models, posing challenges in efficiently updating multiple models and adapting to new models. To address this, we propose OnceEdit, a novel ensemble-based approach that employs a plug-in model as the editing module, enabling stable knowledge updates across multiple models. Building on the model ensemble, OnceEdit introduces two key mechanisms to enhance its effectiveness. First, we introduce a dynamic weight mechanism through a weight token for distinguishing between edit-related and non-edit-related instances, ensuring the appropriate utilization of knowledge from integrated models. Second, we incorporate an ensemble enhancement mechanism to mitigate the excessive reliance on the central model inherent in the model ensemble technique, making it more suitable for knowledge editing. Extensive experiments on diverse LLMs demonstrate that OnceEdit consistently outperforms existing methods while achieving superior editing efficiency. Further analysis confirms its adaptability and stability in multi-model editing scenarios.

Neuron-Level Sequential Editing for Large Language Models

Houcheng Jiang, Junfeng Fang, Tianyu Zhang, Baolong Bi, An Zhang, Ruipeng Wang, Tao Liang, Xiang Wang

This work explores sequential model editing in large language models (LLMs), a critical task that involves modifying internal knowledge within LLMs continuously through multi-round editing, each incorporating updates or corrections to adjust the model's outputs without the need for costly retraining. Existing model editing methods, especially those that alter model parameters, typically focus on single-round editing and often face significant challenges in sequential model editing-most notably issues of model forgetting and failure. To address these challenges, we introduce a new model editing method, namely \textbf{N}euron-level \textbf{S}equential \textbf{E}diting (NSE), tailored for supporting sequential model editing. Specifically, we optimize the target layer's hidden states using the model's original weights to prevent model failure. Furthermore, we iteratively select neurons in multiple layers for editing based on their activation values to mitigate model forgetting. Our empirical experiments demonstrate that NSE significantly outperforms current modifying parameters model editing methods, marking a substantial advancement in the field of sequential model editing. Our code is released on .

Graphically Speaking: Unmasking Abuse in Social Media with Conversation Insights

Célia Nouri, Chloé Clavel, Jean-Philippe Cointet

Detecting abusive language in social media conversations poses significant challenges, as identifying abusiveness often depends on the conversational context, characterized by the content and topology of preceding comments. Traditional Abusive Language Detection (ALD) models often overlook this context, which can lead to unreliable performance metrics. Recent Natural Language Processing (NLP) approaches that incorporate conversational context often rely on limited or overly simplified representations of this context, leading to inconsistent and sometimes inconclusive results. In this paper, we propose a novel approach that utilizes graph neural networks (GNNs) to model social media conversations as graphs, where nodes represent comments, and edges capture reply structures. We systematically investigate various graph representations and context windows to identify the optimal configurations for ALD. Our GNN model outperforms both context-agnostic baselines and linear context-aware methods, achieving significant improvements in F1 scores. These findings demonstrate the critical role of structured conversational context and establish GNNs as a robust framework for advancing context-aware ALD.

CodeTool: Enhancing Programmatic Tool Invocation of LLMs via Process Supervision

YifeiLu, Fanghua Ye, Jian Li, Qiang Gao, Cheng Liu, Haibo Luo, nan du, Xiaolong Li, Feiliang Ren

Tool invocation significantly enhances the capabilities of Large Language Models (LLMs), yet challenges persist, particularly in complex task scenarios. Current methods, such as instruction-enhanced reasoning and supervised fine-tuning, often result in unnecessarily long reasoning paths and face difficulties in verifying the correctness of intermediate steps. In this paper, we propose CodeTool, a novel framework for stepwise code generation that improves LLM tool invocation by leveraging the concise and easily verifiable nature of code. CodeTool incorporates two distinct process rewards: the On-the-spot Reward, which provides immediate feedback on the accuracy of each tool invocation, and the Latent Reward, which assesses the contribution of each step toward overall task completion. By maximizing the cumulative reward of the On-the-spot and Latend Rewards at each step, LLMs are guided to follow efficient and accurate reasoning paths. Extensive experiments on StableToolBench and RestBench-TMDB demonstrate the superiority of CodeTool over existing approaches.

Defense Against Prompt Injection Attack by Leveraging Attack Techniques

Yulin Chen, Haoran Li, Zihao Zheng, Dekai Wu, Yangqiu Song, Bryan Hooi

With the advancement of technology, large language models (LLMs) have achieved remarkable performance across various natural language processing (NLP) tasks, powering LLM-integrated applications like Microsoft Copilot. However, as LLMs continue to evolve, new vulnerabilities, especially prompt injection attacks arise. These attacks trick LLMs into deviating from the original input instructions and executing the attacker's instructions injected in data content, such as retrieved results. Recent attack methods leverage LLMs' instruction-following abilities and their inabilities to distinguish instructions injected in the data content, and achieve a high attack success rate (ASR). When comparing the attack and defense methods, we interestingly find that they share similar design goals, of inducing the model to ignore unwanted instructions and instead to execute wanted instructions. Therefore, we raise an intuitive question: *Could these attack techniques be utilized for defensive purposes?* In this paper, we invert the intention of prompt injection methods to develop novel defense methods based on previous training-free attack methods, by repeating the attack process but with the original input instruction rather than the injected instruction. Our comprehensive experiments demonstrate that our defense techniques outperform existing defense approaches, achieving state-of-the-art results.

Tunable LLM-based Proactive Recommendation Agent

Mingze Wang, Chongming Gao, Wenjie Wang, Yangyang Li, Fuli Feng

Recommender systems are indispensable on various digital platforms. However, traditional methods often reinforce existing user interests, which leads to echo chambers and limits diversity. Proactive Recommendation Systems (PRS) aim to address this issue by cultivating users' latent interests through multi-step recommendations. Despite advancements, challenges persist particularly in optimizing long-term rewards and adapting to real-time user feedback. In this study, we propose an LLM-based Actor-Critic Agent framework to enhance PRS. This framework utilizes the LLM-based agent to adjust recommendations in real time based on feedback and employs agent-tuning methods to optimize long-term rewards using three proposed reward functions. Extensive experiments validate the significant superiority of this framework over existing methods by optimizing long-term rewards and dynamically evolving with user feedback.

AutoMixer: Checkpoint Artifacts as Automatic Data Mixers

Ernie Chang, Yang Li, Patrick Huber, Vish Vogeti, David Kant, Yangyang Shi, Vikas Chandra

In language model training, it is desirable to equip models with capabilities from various tasks. However, it is not clear how to directly obtain the right data mixtures for these capabilities as the relationship between data and tasks is difficult to be modeled. In this work, we observe that checkpoint models exhibit emerging capabilities at different points in the training trajectory. Often, the training process saves checkpoints as artifacts that are under-utilized as a source of in-training data signals. We identify these artifact models based on their respective capabilities on the benchmarks and leverage them as data mixers by using their aggregated first-order influence approximation over source data. We demonstrated on eight reasoning benchmarks that the proposed framework shows significant improvements in the pretraining setting, with accuracy increases of up to 1.93%. Overall, this demonstrates the potential of checkpoint models to enhance data quality and optimize data mixtures.

Consistent Client Simulation for Motivational Interviewing-based Counseling

Yizhe Yang, Palakorn Achananuparp, Heyan Huang, Jing Jiang, Nicholas Gabriel Lim, Cameron Tan Shi Ern, Phey Ling KIT, Jenny Giam Xiuhui, John Pinto, Ee-Peng Lim

Simulating human clients in mental health counseling is crucial for training and evaluating counselors (both human or simulated) in a scalable manner. Nevertheless, past research on client simulation did not focus on complex conversation tasks such as mental health counseling. In these tasks, the challenge is to ensure that the client's actions (i.e., interactions with the counselor) are consistent with with its stipulated profiles and negative behavior settings. In this paper, we propose a novel framework that supports consistent client simulation for mental health counseling. Our framework tracks the mental state of a simulated client, controls its state transitions, and generates for each state behaviors consistent with the client's motivation, beliefs, preferred plan to change, and receptivity. By varying the client profile and receptivity, we demonstrate that consistent simulated clients for different counseling scenarios can be effectively created. Both our automatic and expert evaluations on the generated counseling sessions also show that our client simulation method achieves higher consistency than previous methods.

Structural Reasoning Improves Molecular Understanding of LLM

Yunhui Jang, Jaehyung Kim, Sungsoo Ahn

Recently, large language models (LLMs) have shown significant progress, approaching human perception levels. In this work, we demonstrate that despite these advances, LLMs still struggle to reason using molecular structural information. This gap is critical because many molecular properties, including functional groups, depend heavily on such structural details. To address this limitation, we propose an approach that sketches molecular structures for reasoning. Specifically, we introduce Molecular Structural Reasoning (MSR) framework to enhance the understanding of LLMs by explicitly incorporating the key structural features. We present two frameworks for scenarios where the target molecule is known or unknown. We verify that our MSR improves molecular understanding through extensive experiments.

CAMI: A Counselor Agent Supporting Motivational Interviewing through State Inference and Topic Exploration

Yizhe Yang, Palakorn Achananuparp, Heyan Huang, Jing Jiang, Phey Ling KIT, Nicholas Gabriel Lim, Cameron Tan Shi Ern, Ee-Peng Lim

Conversational counselor agents have become essential tools for addressing the rising demand for scalable and accessible mental health support. This paper introduces CAMI, a novel automated counselor agent grounded in Motivational Interviewing (MI) -- a client-centered counseling approach designed to address ambivalence and facilitate behavior change. CAMI employs a novel STAR framework, consisting of client's state inference, motivation topic exploration, and response generation modules, leveraging large language models (LLMs). These components work together to evoke change talk, aligning with MI principles and improving counseling outcomes for diverse clients. We evaluate CAMI's performance through both automated and expert evaluations, utilizing simulated clients to assess MI skill competency, client's state inference accuracy, topic exploration proficiency, and overall counseling success. Results show that CAMI not only outperforms several state-of-the-art methods but also shows more realistic counselor-like behavior. Additionally, our ablation study underscores the critical roles of state inference and topic exploration in achieving this performance.

Detecting LLM-Generated Korean Text through Linguistic Feature Analysis

Shinwoo Park, Shubin Kim, Do-Kyung Kim, Yo-Sub Han

The rapid advancement of large language models (LLMs) increases the difficulty of distinguishing between human-written and LLM-generated text. Detecting LLM-generated text is crucial for upholding academic integrity, preventing plagiarism, protecting copyrights, and ensuring ethical research practices. Most prior studies on detecting LLM-generated text focus primarily on English text. However, languages with distinct morphological and syntactic characteristics require specialized detection approaches. Their unique structures and usage patterns hinder the direct application of methods primarily designed for English. Among such languages, we focus on Korean, which has relatively flexible spacing rules, a rich morphological system, and less frequent comma usage compared to English. We introduce KatFish, the first benchmark dataset for detecting LLM-generated Korean text. The dataset consists of text written by humans and generated by four LLMs across three genres. By examining spacing patterns, part-of-speech diversity, and comma usage, we illuminate the linguistic differences between human-written and LLM-generated Korean text. Building on these observations, we propose KatFishNet, a detection method specifically designed for the Korean language. KatFishNet achieves an average of 19.78% higher AUC-ROC compared to the best-performing existing detection method. Our code and data are available at https://github.com/Shinwoo-Park/katfishnet.

Uncovering the Impact of Chain-of-Thought Reasoning for Direct Preference Optimization: Lessons from Text-to-SQL

Hanbing Liu, Haoyang Li, Xiaokang Zhang, Ruotong Chen, Haiyong Xu, Tian Tian, Qi Qi, Jing Zhang

Direct Preference Optimization (DPO) has proven effective in complex reasoning tasks like math word problems and code generation. However, when applied to Text-to-SQL datasets, it often fails to improve performance and can even degrade it. Our investigation reveals the root cause: unlike math and code tasks, which naturally integrate Chain-of-Thought (CoT) reasoning with DPO, Text-to-SQL datasets typically include only final answers (gold SQL queries) without detailed CoT solutions. By augmenting Text-to-SQL datasets with synthetic CoT solutions, we achieve, for the first time, consistent and significant performance improvements using DPO. Our analysis shows that CoT reasoning is crucial for unlocking DPO’s potential, as it mitigates reward hacking, strengthens discriminative capabilities, and improves scalability. These findings offer valuable insights for building more robust Text-to-SQL models. To support further research, we publicly release the code and CoT-enhanced datasets: https://github.com/RUCKBReasoning/DPO_Text2SQL.

CogniBench: A Legal-inspired Framework and Dataset for Assessing Cognitive Faithfulness of Large Language Models

Xiaqiang Tang, Jian Li, Keyu Hu, nan du, Xiaolong Li, Xi Zhang, Weigao Sun, Sihong Xie

Faithfulness hallucinations are claims generated by a Large Language Model (LLM) not supported by contexts provided to the LLM. Lacking assessment standards, existing benchmarks focus on "factual statements" that rephrase source materials while overlooking "cognitive statements" that involve making inferences from the given context. Consequently, evaluating and detecting the hallucination of cognitive statements remains challenging. Inspired by how evidence is assessed in the legal domain, we design a rigorous framework to assess different levels of faithfulness of cognitive statements and introduce the CogniBench dataset where we reveal insightful statistics. To keep pace with rapidly evolving LLMs, we further develop an automatic annotation pipeline that scales easily across different models. This results in a large-scale CogniBench-L dataset, which facilitates training accurate detectors for both factual and cognitive hallucinations. We release our model and datasets at: https://github.com/FUTUREEEEEE/CogniBench

Don't Erase, Inform! Detecting and Contextualizing Harmful Language in Cultural Heritage Collections

Orfeas Menis Mastromichalakis, Jason Liartis, Kristina Rose, Antoine Isaac, Giorgos Stamou

Cultural Heritage (CH) data hold invaluable knowledge, reflecting the history, traditions, and identities of societies, and shaping our understanding of the past and present. However, many CH collections contain outdated or offensive descriptions that reflect historical biases. CH Institutions (CHIs) face significant challenges in curating these data due to the vast scale and complexity of the task. To address this, we develop an AI-powered tool that detects offensive terms in CH metadata and provides contextual insights into their historical background and contemporary perception. We leverage a multilingual vocabulary co-created with marginalized communities, researchers, and CH professionals, along with traditional NLP techniques and Large Language Models (LLMs). Available as a standalone web app and integrated with major CH platforms, the tool has processed over 7.9 million records, contextualizing the contentious terms detected in their metadata. Rather than erasing these terms, our approach seeks to inform, making biases visible and providing actionable insights for creating more inclusive and accessible CH collections.

HACo-Det: A Study Towards Fine-Grained Machine-Generated Text Detection under Human-AI Coauthoring

Zhixiong Su, Yichen Wang, Herun Wan, Zhaohan Zhang, Minnan Luo

The misuse of large language models (LLMs) poses potential risks, motivating the development of machine-generated text (MGT) detection. Existing literature primarily concentrates on binary, document-level detection, thereby neglecting texts that are composed jointly by human and LLM contributions. Hence, this paper explores the possibility of fine-grained MGT detection under human-AI coauthoring. We suggest fine-grained detectors can pave pathways toward coauthored text detection with a numeric AI ratio. Specifically, we propose a dataset, HACo-Det, which produces human-AI coauthored texts via an automatic pipeline with word-level attribution labels. We retrofit seven prevailing document-level detectors to generalize them to word-level detection. Then we evaluate these detectors on HACo-Det on both word- and sentence-level detection tasks. Empirical results show that metric-based methods struggle to conduct fine-grained detection with a 0.462 average F1 score, while finetuned models show superior performance and better generalization across domains. However, we argue that fine-grained co-authored text detection is far from solved. We further analyze factors influencing performance, e.g., context window, and highlight the limitations of current methods, pointing to potential avenues for improvement.

Automatic Evaluation for Text-to-image Generation: Task-decomposed Framework, Distilled Training, and Meta-evaluation Benchmark

Rong-Cheng Tu, Zi-Ao Ma, Tian Lan, Yuehao Zhao, Heyan Huang, Xian-Ling Mao

Driven by the remarkable progress in diffusion models, text-to-image generation has achieved substantial advancements, underscoring the urgent need for robust automatic quality assessment. This task is inherently complex, requiring evaluations that range from object presence and attribute correctness to relational consistency and visual fidelity. Consequently, current state-of-the-art MLLM-based approaches often rely on powerful commercial models such as GPT-4o, which offer superior reasoning and instruction-following capabilities but are not universally accessible. In contrast, while open-source MLLMs demonstrate promising skills in vision and language understanding, they underperform in comprehensive image quality assessment. To address these challenges, we propose a task decomposition evaluation framework based on GPT-4o to automatically construct a specialized training dataset, breaking down the multifaceted evaluation process into simpler sub-tasks and thus reducing learning complexity. Building on this dataset, we design novel training strategies to distill GPT-4o's evaluation capabilities into a 7\text{B} open-source MLLM, MiniCPM-V-2.6, enabling it to better follow instructions across diverse assessment criteria. Furthermore, to reliably and comprehensively assess prior works and our proposed model, we manually annotate a meta-evaluation benchmark that includes chain-of-thought explanations alongside quality scores for generated images. Experimental results demonstrate that our distilled open-source MLLM significantly outperforms the current state-of-the-art GPT-4o-base baseline, VIEScore, with over 4.6\% improvement in Spearman and Kendall correlations with human judgments.

TableLoRA: Low-rank Adaptation on Table Structure Understanding for Large Language Models

Xinyi He, Yihao Liu, Mengyu Zhou, Yeye He, Haoyu Dong, Shi Han, Zejian Yuan, Dongmei Zhang

Tabular data are crucial in many fields and their understanding by large language models (LLMs) under high parameter efficiency paradigm is important. However, directly applying parameter-efficient fine-tuning (PEFT) techniques to tabular tasks presents significant challenges, particularly in terms of better table serialization and the representation of two-dimensional structured information within a one-dimensional sequence. To address this, we propose TableLoRA, a module designed to improve LLMs' understanding of table structure during PEFT. It incorporates special tokens for serializing tables with special token encoder and uses 2D LoRA to encode low-rank information on cell positions. Experiments on four tabular-related datasets demonstrate that TableLoRA consistently outperforms vanilla LoRA and surpasses various table encoding methods tested in control experiments. These findings reveal that TableLoRA, as a table-specific LoRA, enhances the ability of LLMs to process tabular data effectively, especially in low-parameter settings, demonstrating its potential as a robust solution for handling table-related tasks.

S²R: Teaching LLMs to Self-verify and Self-correct via Reinforcement Learning

Ruotian Ma, Peisong Wang, Cheng Liu, Xingyan Liu, Jiaqi Chen, Bang Zhang, Xin Zhou, nan du, Jia Li

Recent studies have demonstrated the effectiveness of LLM test-time scaling. However, existing approaches to incentivize LLMs' deep thinking abilities generally require large-scale data or significant training efforts. Meanwhile, it remains unclear how to improve the thinking abilities of less powerful base models. In this work, we introduce S^2R, an efficient framework that enhances LLM reasoning by teaching models to self-verify and self-correct during inference. Specifically, we first initialize LLMs with iterative self-verification and self-correction behaviors through supervised fine-tuning on carefully curated data. The self-verification and self-correction skills are then further strengthened by outcome-level and process-level reinforcement learning with minimized resource requirements. Our results demonstrate that, with only 3.1k behavior initialization samples, Qwen2.5-math-7B achieves an accuracy improvement from 51.0% to 81.6%, outperforming models trained on an equivalent amount of long-CoT distilled data. We also discuss the effect of different RL strategies on enhancing LLMs' deep reasoning. Extensive experiments and analysis based on three base models across both in-domain and out-of-domain benchmarks validate the effectiveness of S^2R.

Retrieval-Augmented Fine-Tuning With Preference Optimization For Visual Program Generation

Deokhyung Kang, Jeonghun Cho, Yejin Jeon, Sunbin Jang, Minsub Lee, JAWOON CHO, Gary Lee

Visual programming languages (VPLs) allow users to create programs through graphical interfaces, which results in easier accessibility and their widespread usage in various domains. To further enhance this accessibility, recent research has focused on generating VPL code from user instructions using large language models (LLMs). Specifically, by employing prompting-based methods, these studies have shown promising results. Nevertheless, such approaches can be less effective for industrial VPLs such as Ladder Diagram (LD). LD is a pivotal language used in industrial automation processes and involves extensive domain-specific configurations, which are difficult to capture in a single prompt. In this work, we demonstrate that training-based methods outperform prompting-based methods for LD generation accuracy, even with smaller backbone models. Building on these findings, we propose a two-stage training strategy to further enhance VPL generation. First, we employ retrieval-augmented fine-tuning to leverage the repetitive use of subroutines commonly seen in industrial VPLs. Second, we apply direct preference optimization (DPO) to further guide the model toward accurate outputs, using systematically generated preference pairs through graph editing operations. Extensive experiments on real-world LD data demonstrate that our approach improves program-level accuracy by over 10\% compared to supervised fine-tuning, which highlights its potential to advance industrial automation.

CENTAUR: Bridging the Impossible Trinity of Privacy, Efficiency, and Performance in Privacy-Preserving Transformer Inference

Jinglong Luo, Guanzhong Chen, Yehong Zhang, SHIYU LIU, Hui Wang, Yue Yu, Xun Zhou, Yuan Qi, Zenglin Xu

With the growing deployment of pre-trained models like Transformers on cloud platforms, privacy concerns about model parameters and inference data are intensifying. Existing Privacy-Preserving Transformer Inference (PPTI) frameworks face the "impossible trinity" of balancing privacy, efficiency, and performance: Secure Multi-Party Computation (SMPC)-based approaches ensure strong privacy but suffer from high computational overhead and performance losses; Conversely, permutation-based methods achieve near-plaintext efficiency and accuracy but compromise privacy by exposing sensitive model parameters and intermediate results. Bridging this gap with a single approach presents substantial challenges, motivating the introduction of CENTAUR, a groundbreaking PPTI framework that seamlessly integrates random permutations and SMPC to address the ``impossible trinity". By designing efficient PPTI algorithms tailored to the structural properties of Transformer models, CENTAUR achieves an unprecedented balance among privacy, efficiency, and performance. Our experiments demonstrate CENTAUR's ability to resist diverse data reconstruction attacks, achieve plaintext-level inference accuracy, and boost inference speed by 5.0\sim30.4 times, unlocking new possibilities for secure and efficient AI deployment.

MT-RAIG: Novel Benchmark and Evaluation Framework for Retrieval-Augmented Insight Generation over Multiple Tables

Kwangwook Seo, Donguk Kwon, Dongha Lee

Recent advancements in table-based reasoning have expanded beyond factoid-level QA to address insight-level tasks, where systems should synthesize implicit knowledge in the table to provide explainable analyses. Although effective, existing studies remain confined to scenarios where a single gold table is given alongside the user query, failing to address cases where users seek comprehensive insights from multiple unknown tables. To bridge these gaps, we propose MT-RAIG Bench, design to evaluate systems on Retrieval-Augmented Insight Generation over Mulitple-Tables. Additionally, to tackle the suboptimality of existing automatic evaluation methods in the table domain, we further introduce a fine-grained evaluation framework MT-RAIG Eval, which achieves better alignment with human quality judgments on the generated insights. We conduct extensive experiments and reveal that even frontier LLMs still struggle with complex multi-table reasoning, establishing our MT-RAIG Bench as a challenging testbed for future research.

Towards Widening The Distillation Bottleneck for Reasoning Models

Huifeng Yin, Yu Zhao, Minghao Wu, Xuanfan Ni, Bo Zeng, huaiyu.wh, Tianqi Shi, Liangying Shao, Chenyang Lyu, Longyue Wang, Weihua Luo, Kaifu Zhang

Large Reasoning Models (LRMs) such as OpenAI o1 and DeepSeek-R1 have shown remarkable reasoning capabilities by scaling test-time compute and generating long Chain-of-Thought (CoT). Distillation post-training on LRMs-generated data is a straightforward yet effective method to enhance the reasoning abilities of smaller models, but faces a critical bottleneck: we found that distilled long CoT data poses learning difficulty for small models and leads to the inheritance of biases (i.e., formalistic long-time thinking) when using Supervised Fine-tuning (SFT) and Reinforcement Learning (RL) methods. To alleviate this bottleneck, we propose constructing data from scratch using Monte Carlo Tree Search (MCTS). We then exploit a set of CoT-aware approaches, including Thoughts Length Balance, Fine-grained DPO, and Joint Post-training Objective, to enhance SFT and RL on the MCTS data. We conducted evaluation on various benchmarks such as math (GSM8K, MATH, AIME). instruction-following (Multi-IF) and planning (Blocksworld), results demonstrate our CoT-aware approaches substantially improve the reasoning performance of distilled models compared to standard distilled models via reducing the hallucinations in long-time thinking.

Can We Further Elicit Reasoning in LLMs? Critic-Guided Planning with Retrieval-Augmentation for Solving Challenging Tasks

Xingxuan Li, Weiwen Xu, Ruochen Zhao, Fangkai Jiao, Shafiq Joty, Lidong Bing

Large language models excel at problem-solving but often struggle with complex reasoning and factual accuracy. While chain-of-thought and retrieval-augmented generation help break down problems and retrieve knowledge, they still falter on challenging tasks like competitive programming due to frequent reasoning errors and irrelevant retrieval. To address this, we introduce Critic-guided planning with Retrieval-augmentation, CR-Planner, a novel framework that leverages fine-tuned critic models to guide both reasoning and retrieval processes through planning. CR-Planner iteratively selects and executes sub-goals, guided by critic models. A sub-goal critic identifies promising sub-goals from reasoning, query generation, and retrieval, while an execution critic evaluates outputs of sub-goal executions. We employ Monte Carlo Tree Search to collect data for critic training, allowing systematic exploration of action sequences and effective navigation toward the final answer. We evaluate CR-Planner on challenging domain-knowledge-intensive and reasoning-heavy tasks, including competitive programming, theorem-driven math reasoning, and complex domain retrieval problems. It significantly outperforms baselines, demonstrating effectiveness in both reasoning and retrieval.

RADAR: Enhancing Radiology Report Generation with Supplementary Knowledge Injection

Wenjun Hou, Yi Cheng, Kaishuai Xu, Heng Li, Yan Hu, Wenjie Li, Jiang Liu

Large language models (LLMs) have demonstrated remarkable capabilities in various domains, including radiology report generation. Previous approaches have attempted to utilize multimodal LLMs for this task, enhancing their performance through the integration of domain-specific knowledge retrieval. However, these approaches often overlook the knowledge already embedded within the LLMs, leading to redundant information integration. To address this limitation, we propose Radar, a framework for enhancing radiology report generation with supplementary knowledge injection. Radar improves report generation by systematically leveraging both the internal knowledge of an LLM and externally retrieved information. Specifically, it first extracts the model's acquired knowledge that aligns with expert image-based classification outputs. It then retrieves relevant supplementary knowledge to further enrich this information. Finally, by aggregating both sources, Radar generates more accurate and informative radiology reports. Extensive experiments on MIMIC-CXR, CheXpert-Plus, and IU X-ray demonstrate that our model outperforms state-of-the-art LLMs in both language quality and clinical accuracy

WAVE: Building Multimodal Web Agents via Iterative Real-World Exploration, Feedback and Optimization

Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Hongming Zhang, Tianqing Fang, Zhenzhong Lan, Dong Yu

The advancement of foundation models has laid the groundwork for building autonomous agents for complex tasks such as web navigation. Recent efforts have also tried to equip the agent with the ability to explore environments and continuously improve over time. However, existing works only focused on building text-only agents in synthetic environments where the reward signals are clearly defined. Such agents can hardly generalize to realistic settings that require multimodal perception ability and provide no ground-truth signal. In this paper, we introduce an innovative multimodal web agent that can autonomously conduct real-world exploration and improve itself. We first train the base model with imitation learning to gain the basic abilities. We then let the agent explore the open web and collect feedback on its trajectories. After that, it further improves its policy by learning from well-performing trajectories judged by another general-purpose model. This exploration-feedback-optimization cycle can continue for several iterations. Experimental results show that our web agent successfully improves itself after each iteration, demonstrating strong performance across multiple test sets. We will release our code and model to encourage future research in this field.

Mixture of Small and Large Models for Chinese Spelling Check

Ziheng Qiao, Houquan Zhou, Zhenghua Li

In the era of large language models (LLMs), the Chinese Spelling Check (CSC) task has seen various LLM methods developed, yet their performance remains unsatisfactory. In contrast, fine-tuned BERT-based models, relying on high-quality in-domain data, show excellent performance but suffer from edit pattern overfitting. This paper proposes a novel dynamic mixture approach that effectively combines the probability distributions of small models and LLMs during the beam search decoding phase, achieving a balanced enhancement of precise corrections from small models and the fluency of LLMs. This approach also eliminates the need for fine-tuning LLMs, saving significant time and resources, and facilitating domain adaptation. Comprehensive experiments demonstrate that our mixture approach significantly boosts error correction capabilities, achieving state-of-the-art results across multiple datasets. Our code is available at https://github.com/zhqiao-nlp/MSLLM.

Medical Graph RAG: Evidence-based Medical Large Language Model via Graph Retrieval-Augmented Generation

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, Yueming Jin, Vicente Grau

We introduce MedGraphRAG, a novel graph-based Retrieval-Augmented Generation (RAG) framework designed to enhance LLMs in generating evidence-based medical responses, improving safety and reliability with private medical data. We introduce Triple Graph Construction and U-Retrieval to enhance GraphRAG, enabling holistic insights and evidence-based response generation for medical applications. Specifically, we connect user documents to credible medical sources and integrate Top-down Precise Retrieval with Bottom-up Response Refinement for balanced context awareness and precise indexing. Validated on 9 medical Q\&A benchmarks, 2 health fact-checking datasets, and a long-form generation test set, MedGraphRAG outperforms state-of-the-art models while ensuring credible sourcing. Our code is publicly available.

A3: Automatic Alignment Framework for Attributed Text Generation

Yue Wang, Haoke Zhang, Juntao Li, Jinxiong Chang, Min Zhang

Attributed text generation aims to enhance the reliability of content generated from large language models by providing citations for each claim, which thereby enables users to easily verify the correctness of the responses. However, the scarcity of high-quality training samples presents a significant challenge in aligning large language models to generate texts with citations, revealing considerable room for improvement in existing attribution systems. Besides, existing approaches of aligning large language models to follow user instructions can lead to an undue emphasis on irrelevant documents, which in turn reduces the quality of responses. To address the above problems, we propose \textbf{A}utomatic \textbf{A}lignment Framework for \textbf{A}ttributed Text Generation(\textbf{\mathcal{A}^3}), a novel framework designed to automatically generate high-quality attributed query-response pairs for both supervised fine-tuning and preference optimization stages without human annotation. With the help of \textbf{\mathcal{A}^3}, Mistral-7B can achieve a citation recall of \textbf{84.4} and a precision of \textbf{87.0} precision on ASQA, which notably surpasses GPT-4's citation recall of \textbf{73.0} and precision of \textbf{76.5}.

DeepReview: Improving LLM-based Paper Review with Human-like Deep Thinking Process

Minjun Zhu, Yixuan Weng, Linyi Yang, Yue Zhang

Large Language Models (LLMs) are increasingly utilized in scientific research assessment, particularly in automated paper review. However, existing LLM-based review systems face significant challenges, including limited domain expertise, hallucinated reasoning, and a lack of structured evaluation. To address these limitations, we introduce DeepReview, a multi-stage framework designed to emulate expert reviewers by incorporating structured analysis, literature retrieval, and evidence-based argumentation. Using DeepReview-13K, a curated dataset with structured annotations, we train DeepReviewer-14B, which outperforms CycleReviewer-70B with fewer tokens. In its best mode, DeepReviewer-14B achieves win rates of 88.21% and 80.20% against GPT-o1 and DeepSeek-R1 in evaluations. Our work sets a new benchmark for LLM-based paper review, with all resources publicly available.

Beyond True or False: Retrieval-Augmented Hierarchical Analysis of Nuanced Claims

Priyanka Kargupta, Runchu Tian, Jiawei Han

Claims made by individuals or entities are oftentimes nuanced and cannot be clearly labeled as entirely "true" or "false"---as is frequently the case with scientific and political claims. However, a claim (e.g., "vaccine A is better than vaccine B") can be dissected into its integral aspects and sub-aspects (e.g., efficacy, safety, distribution), which are individually easier to validate. This enables a more comprehensive, structured response that provides a well-rounded perspective on a given problem while also allowing the reader to prioritize specific angles of interest within the claim (e.g., safety towards children). Thus, we propose ClaimSpect, a retrieval-augmented generation-based framework for automatically constructing a hierarchy of aspects typically considered when addressing a claim and enriching them with corpus-specific perspectives. This structure hierarchically partitions an input corpus to retrieve relevant segments, which assist in discovering new sub-aspects. Moreover, these segments enable the discovery of varying perspectives towards an aspect of the claim (e.g., support, neutral, or oppose) and their respective prevalence (e.g., "how many biomedical papers believe vaccine A is more transportable than B?"). We apply ClaimSpect to a wide variety of real-world scientific and political claims featured in our constructed dataset, showcasing its robustness and accuracy in deconstructing a nuanced claim and representing perspectives within a corpus. Through real-world case studies and human evaluation, we validate its effectiveness over multiple baselines.

Theorem Prover as a Judge for Synthetic Data Generation

Joshua Ong Jun Leang, Giwon Hong, Wenda Li, Shay B Cohen

The demand for synthetic data in mathematical reasoning has increased due to its potential to enhance the mathematical capabilities of large language models (LLMs). However, ensuring the validity of intermediate reasoning steps remains a significant challenge, affecting data quality. While formal verification via theorem provers effectively validates LLM reasoning, the autoformalisation of mathematical proofs remains error-prone. In response, we introduce *iterative autoformalisation*, an approach that iteratively refines theorem prover formalisation to mitigate errors, thereby increasing the execution rate on the Lean prover from 60% to 87%. Building upon that, we introduce *Theorem Prover as a Judge (TP-as-a-Judge)*, a method that employs theorem prover formalisation to rigorously assess LLM intermediate reasoning, effectively integrating autoformalisation with synthetic data generation. Finally, we present *Reinforcement Learning from Theorem Prover Feedback (RLTPF),* a framework that replaces human annotation with theorem prover feedback in Reinforcement Learning from Human Feedback (RLHF). Across multiple LLMs, applying *TP-as-a-Judge* and *RLTPF* improves benchmarks with only 3,508 samples, achieving 5.56% accuracy gain on Mistral-7B for MultiArith, 6.00% on Llama-2-7B for SVAMP, and 3.55% on Llama-3.1-8B for AQUA.

VISA: Retrieval Augmented Generation with Visual Source Attribution

Xueguang Ma, Shengyao Zhuang, Bevan Koopman, Guido Zuccon, Wenhu Chen, Jimmy Lin

Generation with source attribution is important for enhancing the verifiability of retrieval-augmented generation (RAG) systems. However, existing approaches in RAG primarily link generated content to document-level references, making it challenging for users to locate evidence among multiple content-rich retrieved documents. To address this challenge, we propose Retrieval-Augmented Generation with Visual Source Attribution (VISA), a novel approach that combines answer generation with visual source attribution. Leveraging large vision-language models (VLMs), VISA identifies the evidence and highlights the exact regions that support the generated answers with bounding boxes in the retrieved document screenshots. To evaluate its effectiveness, we curated two datasets: Wiki-VISA, based on crawled Wikipedia webpage screenshots, and Paper-VISA, derived from PubLayNet and tailored to the medical domain. Experimental results demonstrate the effectiveness of VISA for visual source attribution on documents' original look, as well as highlighting the challenges for improvement.

IRIS: Interpretable Retrieval-Augmented Classification for Long Interspersed Document Sequences

Fengnan Li, Elliot D. Hill, JIANG SHU, Jiaxin Gao, Matthew M. Engelhard

Transformer-based models have achieved state-of-the-art performance in document classification but struggle with long-text processing due to the quadratic computational complexity in the self-attention module. Existing solutions, such as sparse attention, hierarchical models, and key sentence extraction, partially address the issue but still fall short when the input sequence is exceptionally lengthy. To address this challenge, we propose **IRIS** (**I**nterpretable **R**etrieval-Augmented Classification for long **I**nterspersed Document **S**equences), a novel, lightweight framework that utilizes retrieval to efficiently classify long documents while enhancing interpretability. IRIS segments documents into chunks, stores their embeddings in a vector database, and retrieves those most relevant to a given task using learnable query vectors. A linear attention mechanism then aggregates the retrieved embeddings for classification, allowing the model to process arbitrarily long documents without increasing computational cost and remaining trainable on a single GPU. Our experiments across six datasets show that IRIS achieves comparable performance to baseline models on standard benchmarks, and excels in three clinical note disease risk prediction tasks where documents are extremely long and key information is sparse. Furthermore, IRIS provides global interpretability by revealing a clear summary of key risk factors identified by the model. These findings highlight the potential of IRIS as an efficient and interpretable solution for long-document classification, particularly in healthcare applications where both performance and explainability are crucial.

R2D2: Remembering, Replaying and Dynamic Decision Making with a Reflective Agentic Memory

Tenghao Huang, Kinjal Basu, Ibrahim Abdelaziz, Pavan Kapanipathi, Jonathan May, Muhao Chen

The proliferation of web agents necessitates advanced navigation and interaction strategies within complex web environments. Current models often struggle with efficient navigation and action execution due to limited visibility and understanding of web structures. Our proposed R2D2 framework addresses these challenges by integrating two paradigms: Remember and Reflect. The Remember paradigm utilizes a replay buffer that aids agents in reconstructing the web environment dynamically, thus enabling the formulation of a detailed ``map'' of previously visited pages. This helps in reducing navigational errors and optimizing the decision-making process during web interactions. Conversely, the Reflect paradigm allows agents to learn from past mistakes by providing a mechanism for error analysis and strategy refinement, enhancing overall task performance. We evaluate R2D2 using the WEBARENA benchmark, demonstrating significant improvements over existing methods, including a 50% reduction in navigation errors and a threefold increase in task completion rates. Our findings suggest that a combination of memory-enhanced navigation and reflective learning promisingly advances the capabilities of web agents, potentially benefiting various applications such as automated customer service and personal digital assistants.

Meta-Tool: Unleash Open-World Function Calling Capabilities of General-Purpose Large Language Models

Shengqian Qin, Yakun Zhu, Linjie Mu, Shaoting Zhang, Xiaofan Zhang

Large language models (LLMs) have showcased remarkable capabilities as autonomous agents when augmented with external tools. Equipped with fixed tool sets, LLMs struggle with addressing diverse user inquiries in open-world tasks. To evaluate and boost the performance of LLMs in dealing with complex demands in the real-world, we propose open-world function calling, where LLMs need to retrieve suitable tools from a pre-defined external tool library and use retrieved tools to resolve the user's problem. We introduce Meta-Tool, a versatile and plug-and-play tool retrieval system as the access of LLMs to external tool library. Drawing inspiration from the myriad of enhanced approaches associated with Retrieval-Augmented Generation (RAG), Meta-Tool employs a \textit{hypothesize-retrieve-invoke} framework. We further propose Meta-Bench, a comprehensive benchmark for evaluating LLMs in open-world function calling and associated tasks. Meta-Bench encompasses 2,800 dialogues and 7,361 tools, spanning ten distinct scenarios to provide robust and diverse test categories. In conjunction, we present MT-LLaMA, a finetuned version of LLaMA-3.1, which exhibits remarkable performance improvements. Our empirical experiments reveal that Meta-Tool significantly enhances the ability of advanced LLMs to retrieve and leverage the most suitable tools compared to previous tool retrieval methods. Moreover, our fine-tuning enables even smaller-sized LLMs to achieve comparable even exceeding results to GPT-4o. Both the benchmark and the model are made publicly available at https://github.com/qinshengqian/Meta-Tool to foster further research and development in the field.

Less is More: Explainable and Efficient ICD Code Prediction with Clinical Entities

James Douglas, Yidong Gan, Ben Hachey, Jonathan K. Kummerfeld

Clinical coding, assigning standardized codes to medical notes, is critical for epidemiological research, hospital planning, and reimbursement. Neural coding models generally process entire discharge summaries, which are often lengthy and contain information that is not relevant to coding. We propose an approach that combines Named Entity Recognition (NER) and Assertion Classification (AC) to filter for clinically important content before supervised code prediction. On MIMIC-IV, a standard evaluation dataset, our approach achieves near-equivalent performance to a state-of-the-art full-text baseline while using only 22% of the content and reducing training time by over half. Additionally, mapping model attention to complete entity spans yields coherent, clinically meaningful explanations, capturing coding-relevant modifiers such as acuity and laterality. We release a newly annotated NER+AC dataset for MIMIC-IV, designed specifically for ICD coding. Our entity-centric approach lays a foundation for more transparent and cost-effective assisted coding.

DDxTutor: Clinical Reasoning Tutoring System with Differential Diagnosis-Based Structured Reasoning

Qian Wu, Zheyao Gao, Longfei Gou, Qi Dou

Clinical diagnosis education requires students to master both systematic reasoning processes and comprehensive medical knowledge. While recent advances in Large Language Models (LLMs) have enabled various medical educational applications, these systems often provide direct answers that could reduce students' cognitive engagement and lead to fragmented learning. Motivated by these challenges, we propose DDxTutor, a framework that follows differential diagnosis principles to decompose clinical reasoning into teachable components. It consists of a structured reasoning module that analyzes clinical clues and synthesizes diagnostic conclusions, and an interactive dialogue framework that guides students through this process. To enable such tutoring, we construct DDxReasoning, a dataset of 933 clinical cases with fine-grained diagnostic steps verified by doctors. Our experiments demonstrate that fine-tuned LLMs achieve strong performance in generating structured teaching references and conducting interactive diagnostic tutoring dialogues. Human evaluation by medical educators and students validates the framework's potential and effectiveness for clinical diagnosis education. Our project is available at https://github.com/med-air/DDxTutor.

Rethinking Evaluation Metrics for Grammatical Error Correction: Why Use a Different Evaluation Process than Human?

Takumi Goto, Yusuke Sakai, Taro Watanabe

One of the goals of automatic evaluation metrics in grammatical error correction (GEC) is to rank GEC systems such that it matches human preferences. However, current automatic evaluations are based on procedures that diverge from human evaluation. Specifically, human evaluation derives rankings by aggregating sentence-level relative evaluation results, e.g., pairwise comparisons, using a rating algorithm, whereas automatic evaluation averages sentence-level absolute scores to obtain corpus-level scores, which are then sorted to determine rankings. In this study, we propose an aggregation method for existing automatic evaluation metrics which aligns with human evaluation methods to bridge this gap. We conducted experiments using various metrics, including edit-based metrics, -gram based metrics, and sentence-level metrics, and show that resolving the gap improves results for the most of metrics on the SEEDA benchmark.We also found that even BERT-based metrics sometimes outperform the metrics of GPT-4.

Efficient Domain Continual pretraining by Mitigating the Stability Gap

Yiduo Guo, Jie Fu, Huishuai Zhang, Dongyan Zhao

Continual pretraining enables Large Language Models (LLMs) to adapt to specialized domains like medicine and law. However, we observe a consistent phenomenon across different model sizes and domains: a temporary performance drop at the start of the continual pretraining process, followed by a performance recovery phase. To gain a deeper understanding of this issue, we use the stability gap— a concept adapted from the visual domain—which explains this initial drop arises from instability in the model’s general abilities. We validate this hypothesis through a series of experiments. To address this initial instability and enhance LLM performance within a fixed compute budget, we propose a training strategy that mitigates instability by increasing the number of epochs, alongside two data sampling strategies targeting data domain relevance and corpus distribution. We conduct experiments on Llama-family models to validate the effectiveness of our strategies for continual pretraining and instruction tuning in medical and legal domains. Our strategies improve the average medical task performance of the OpenLlama-3B model from 36.2% to 40.7% using only 40% of the original training budget, while also enhancing general task performance without causing forgetting. Furthermore, we aPPLy our strategies to continually pre-train and instruction-tune the Llama-3-8B model. The resulting model, Llama-3-Physician, achieves the best medical performance among open-source models on several benchmarks and rivals GPT-4 on specific tasks. We release our models at https://huggingface.co/YiDuo1999/Llama-3-Physician-8B-Instruct.

Session 12: NLP Applications (Wednesday 30 July, 11:00-12:30, Hall 5X)

Multi-Programming Language Sandbox for LLMs

Shihan Dou, Jiazheng Zhang, Zang Jianxiang, Yunbo Tao, Weikang Zhou, Haoxiang Jia, Shichun Liu, Yuming Yang, Shenxi Wu, Zhiheng Xi, Muling Wu, Rui Zheng, Changze Lv, Limao Xiong, Shaoqing Zhang, Lin Zhang, Wenyu Zhan, Rongxiang Weng, Jingang Wang, Xunliang Cai, Yueming Wu, Ming Wen, Yixin Cao, Tao Gui, Xipeng Qiu, Qi Zhang, Xuanjing Huang

We introduce MPLSandbox, an out-of-the-box multi-programming language sandbox designed to provide unified and comprehensive feedback from compiler and analysis tools for Large Language Models (LLMs). It can automatically identify the programming language of the code, compiling and executing it within an isolated sub-sandbox to ensure safety and stability. In addition, MPLSandbox integrates both traditional and LLM-based code analysis tools, providing a comprehensive analysis of generated code. It also can be effortlessly integrated into the training and deployment of LLMs to improve the quality and correctness of generated code. It also helps researchers streamline their workflows for various LLM-based code-related tasks, reducing the development cost. To validate the effectiveness of MPLSandbox, we conduct extensive experiments by integrating it into several training and deployment scenarios, and employing it to optimize workflows for a wide range of downstream code tasks. Our goal is to enhance researcher productivity on LLM-based code tasks by simplifying and automating workflows through delegation to MPLSandbox.

SPOT: Bridging Natural Language and Geospatial Search for Investigative Journalism

Lynn Khellaf, Ipek Baris Schlicht, Tilman Mirass, Julia Bayer, Tilman Wagner, Ruben Bouwmeester

OpenStreetMap (OSM) is a vital resource for investigative journalists using geolocation verification. However, existing tools such as Overpass Turbo require familiarity with complex query languages, creating barriers for non-technical users. We present SPOT, an open source natural language interface that makes OSM's rich, tag-based geographic data more accessible through intuitive scene descriptions. SPOT interprets user inputs as structured representations of geographic object configurations using fine-tuned Large Language Models (LLMs), with results being displayed in an interactive map interface. While more general geospatial search tasks are conceivable, SPOT is specifically designed for use in investigative journalism, addressing real-world challenges such as hallucinations in model output, inconsistencies in OSM tagging, and the noisy nature of user input. It combines a novel synthetic data pipeline with a semantic bundling system to enable robust, accurate query generation. To our knowledge, SPOT is the first system to achieve reliable natural language access to OSM data at this level of accuracy. By lowering the technical barrier to spatial verification, SPOT contributes a practical tool to the broader efforts to support fact-checking and combat disinformation.

VenusFactory: An Integrated System for Protein Engineering with Data Retrieval and Language Model Fine-Tuning

Yang Tan, Chen Liu, Jingyuan Gao, Wu Banghao, Mingchen Li, Ruilin Wang, Lingrong Zhang, Huiqun Yu, Guisheng Fan, Liang Hong, Bingxin Zhou

Natural language processing (NLP) has significantly influenced scientific domains beyond human language, including protein engineering, where pre-trained protein language models (PLMs) have demonstrated remarkable success. However, interdisciplinary adoption remains limited due to challenges in data collection, task benchmarking, and application. This work presents VenusFactory, a versatile engine that integrates biological data retrieval, standardized task benchmarking, and modular fine-tuning of PLMs. VenusFactory supports both computer science and biology communities with choices of both a command-line execution and a Gradio-based no-code interface, integrating 40+ protein-related datasets and 40+ popular PLMs. All implementations are open-sourced on https://github.com/ai4protein/VenusFactory. A video introduction is available at https://www.youtube.com/watch?v=MT6lPH5kgCc.

DocSpiral: A Platform for Integrated Assistive Document Annotation through Human-in-the-Spiral

Qiang Sun, Sirui Li, Tingting Bi, Du Q. Huynh, Mark Reynolds, Yuanyi Luo, Wei Liu

Acquiring structured data from domain-specific, image-based documents—such as scanned reports—is crucial for many downstream tasks but remains challenging due to document variability. Many of these documents exist as images rather than as machine-readable text, which requires human annotation to train automated extraction systems. We present \textbf{DocSpiral}, the first Human-in-the-Spiral assistive document annotation platform, designed to address the challenge of extracting structured information from domain-specific, image-based document collections. Our spiral design establishes an iterative cycle in which human annotations train models that progressively require less manual intervention. \textbf{DocSpiral} integrates document format normalization, comprehensive annotation interfaces, evaluation metrics dashboard, and API endpoints for the development of AI / ML models into a unified workflow. Experiments demonstrate that our framework reduces annotation time by at least 41\% while showing consistent performance gains across three iterations during model training. By making this annotation platform freely accessible, we aim to lower barriers to AI/ML models development in document processing, facilitating the adoption of large language models in image-based, document-intensive fields such as geoscience and healthcare. The system is freely available at: . The demonstration video is available: .

ADEPT-SQL: A High-performance Text-to-SQL Application for Real-World Enterprise-Level Databases

yongnan chen, Zhuo Chang, Shijia Gu, yuanhang zong, Zhang Mei, Shiyu Wang, HEZIXIANG, Hongzhi Chen, Jin Wei, Bin CUI

This paper presents Adept-SQL, a domain-adapted Text2SQL system that addresses critical deployment challenges in professional fields. While modern LLM-based solutions excel on academic benchmarks, we identify three persistent limitations in industrial application: domain-specific knowledge barriers, the schemas complexity in the real world, and the prohibitive computational costs of large LLMs. Our framework introduces two key innovations: a three-stage grounding mechanism combining dynamic terminology expansion, focused schema alignment, and historical query retrieval; coupled with a hybrid prompting architecture that decomposes SQL generation into schema-aware hinting, term disambiguation, and few-shot example incorporation phases. This approach enables efficient execution using smaller open-source LLMs while maintaining semantic precision. Deployed in petroleum engineering domains, our system achieves 97% execution accuracy on real-world databases, demonstrating 49% absolute improvement over SOTA baselines. We release implementation code to advance research in professional Text2SQL systems.

Live Football Commentary System Providing Background Information

Yuichiro Mori, Chikara Tanaka, Aru Maekawa, Satoshi Kosugi, Tatsuya Ishigaki, Kotaro Funakoshi, Hiroya Takamura, Manabu Okumura

Previous research on sports commentary generation has primarily focused on describing major events in the match. However, real-world commentary often includes comments beyond what is visible in the video content, e.g., ``Florentina has acquired him for 7 million euros." For enhancing the viewing experience with such background information,we developed an audio commentary system for football matches that generates utterances with background information, as well as play-by-play commentary. Our system first extracts visual information, and determines whether it is an appropriate timing to produce an utterance. Then it decides which type of utterance to generate: play-by-play or background information. In the latter case, the system leverages external knowledge through retrieval-augmented generation.

ComfyUI-Copilot: An Intelligent Assistant for Automated Workflow Development

Zhenran Xu, Yangxue, Yiyu Wang, Qingli Hu, Zijiao Wu, Baotian Hu, Longyue Wang, Weihua Luo, Kaifu Zhang

We introduce **ComfyUI-Copilot**, a large language model-powered plugin designed to enhance the usability and efficiency of ComfyUI, an open-source platform for AI-driven art creation. Despite its flexibility and user-friendly interface, ComfyUI can present challenges to newcomers, including limited documentation, model misconfigurations, and the complexity of workflow design. ComfyUI-Copilot addresses these challenges by offering intelligent node and model recommendations, along with automated one-click workflow construction. At its core, the system employs a hierarchical multi-agent framework comprising a central assistant agent for task delegation and specialized worker agents for different usages, supported by our curated ComfyUI knowledge bases to streamline debugging and deployment. We validate the effectiveness of ComfyUI-Copilot through both offline quantitative evaluations and online user feedback, showing that it accurately recommends nodes and accelerates workflow development. Additionally, use cases illustrate that ComfyUI-Copilot lowers entry barriers for beginners and enhances workflow efficiency for experienced users. The ComfyUI-Copilot installation package and a demo video are available at https://github.com/AIDC-AI/ComfyUI-Copilot.

PRAISE: Enhancing Product Descriptions with LLM-Driven Structured Insights

Adnan Qidwai, Srija Mukhopadhyay, Prerana Khatiwada, Dan Roth, Vivek Gupta

Accurate and complete product descriptions are crucial for e-commerce, yet seller-provided information often falls short. Customer reviews offer valuable details but are laborious to sift through manually. We present PRAISE: Product Review Attribute Insight Structuring Engine, a novel system that uses Large Language Models (LLMs) to automatically extract, compare, and structure insights from customer reviews and seller descriptions. PRAISE provides users with an intuitive interface to identify missing, contradictory, or partially matching details between these two sources, presenting the discrepancies in a clear, structured format alongside supporting evidence from reviews. This allows sellers to easily enhance their product listings for clarity and persuasiveness, and buyers to better assess product reliability. Our demonstration showcases PRAISE's workflow, its effectiveness in generating actionable structured insights from unstructured reviews, and its potential to significantly improve the quality and trustworthiness of e-commerce product catalogs.

Session 12: Phonology, Morphology, and Word Segmentation AND Syntax: Tagging, Chunking and Parsing (Wednesday 30 July, 11:00-12:30, Hall 4/5)

BERT-like Models for Slavic Morpheme Segmentation

Dmitry Morozov, Lizaveta Astapenka, Anna Glazkova, Timur Garipov, Olga Lyashevskaya

Automatic morpheme segmentation algorithms are applicable in various tasks, such as building tokenizers and language education. For Slavic languages, the development of such algorithms is complicated by the rich derivational capabilities of these languages. Previous research has shown that, on average, these algorithms have already reached expert-level quality. However, a key unresolved issue is the significant decline in performance when segmenting words containing roots not present in the training data. This problem can be partially addressed by using pre-trained language models to better account for word semantics. In this work, we explored the possibility of fine-tuning BERT-like models for morpheme segmentation using data from Belarusian, Czech, and Russian. We found that for Czech and Russian, our models outperform all previously proposed approaches, achieving word-level accuracy of 92.5-95.1%. For Belarusian, this task was addressed for the first time. The best-performing approach for Belarusian was an ensemble of convolutional neural networks with word-level accuracy of 90.45%.

Comparing LLM-generated and human-authored news text using formal syntactic theory

Olga Zamaraeva, Dan Flickinger, Francis Bond, Carlos Gómez-Rodríguez

This study provides the first comprehensive comparison of New York Times-style text generated by six large language models against real, human-authored NYT writing. The comparison is based on a formal syntactic theory. We use Head-driven Phrase Structure Grammar (HPSG) to analyze the grammatical structure of the texts. We then investigate and illustrate the differences in the distributions of HPSG grammar types, revealing systematic distinctions between human and LLM-generated writing. These findings contribute to a deeper understanding of the syntactic behavior of LLMs as well as humans, within the NYT genre.

Contrastive Learning on LLM Back Generation Treebank for Cross-domain Constituency Parsing

Peiming Guo, Meishan Zhang, jianling li, Min Zhang, Yue Zhang

Cross-domain constituency parsing is still an unsolved challenge in computational linguistics since the available multi-domain constituency treebank is limited. We investigate automatic treebank generation by large language models (LLMs) in this paper. The performance of LLMs on constituency parsing is poor, therefore we propose a novel treebank generation method, LLM back generation, which is similar to the reverse process of constituency parsing. LLM back generation takes the incomplete cross-domain constituency tree with only domain keyword leaf nodes as input and fills the missing words to generate the cross-domain constituency treebank. Besides, we also introduce a span-level contrastive learning pre-training strategy to make full use of the LLM back generation treebank for cross-domain constituency parsing. We verify the effectiveness of our LLM back generation treebank coupled with contrastive learning pre-training on five target domains of MCTB. Experimental results show that our approach achieves state-of-the-art performance on average results compared with various baselines.

Session 12: Question Answering (Wednesday 30 July, 11:00-12:30, Hall 4/5)

Addressing Blind Guessing: Calibration of Selection Bias in Multiple-Choice Question Answering by Video Language Models

Olga Loginova, Oleksandr Bezrukov, Alexey Kravets

Evaluating Video Language Models (VLMs) is a challenging task. Due to its transparency, Multiple-Choice Question Answering (MCQA) is widely used to measure the performance of these models through accuracy. However, existing MCQA benchmarks fail to capture the full reasoning capabilities of VLMs due to selection bias, when models disproportionately favor certain answer options based on positional patterns observed during training. In this work, we conduct a comprehensive empirical analysis of several VLM architectures across major datasets designed to assess complex video-focused reasoning. We identify where the bias is most pronounced and demonstrate to what extent model responses reflect genuine understanding of video content and related questions, as opposed to reliance on arbitrary patterns or superficial cues, such as answer position. By decomposing the MCQA task and adapting fairness bias metrics to VLMs, we introduce a post-processing calibration technique BOLD to balance this bias. Our results show that reducing selection bias improves not only debiasing metrics but also overall model performance, including Accuracy and F1 Mean score. Our method, by suppressing "blind guessing", offers a more cost- and time-effective approach to mitigating selection bias compared to existing techniques. This study represents the first focused investigation of selection bias in video-to-text LLM-powered models.

Fine-Tuning on Diverse Reasoning Chains Drives Within-Inference CoT Refinement in LLMs

Haritz Puerto, Tilek Chubakov, Xiaodan Zhu, Harish Tayyar Madabushi, Iryna Gurevych

Requiring a large language model (LLM) to generate intermediary reasoning steps, known as Chain of Thought (CoT), has been shown to be an effective way of boosting performance. Previous approaches have focused on generating multiple independent CoTs, combining them through ensembling or other post-hoc strategies to enhance reasoning. In this work, we introduce a novel approach where LLMs are fine-tuned to generate a sequence of Diverse Chains of Thought (DCoT) within a single inference step, which is fundamentally different from prior work that primarily operate on parallel CoT generations. DCoT allows LLMs to gain the ability to perform within-inference refinement of reasoning chains without requiring external feedback. Through a rigorous set of experiments spanning a wide range of tasks that require various reasoning types, we show that fine-tuning on DCoT improves performance over the CoT baseline across model families and scales (1.3B to 70B). These improvements are particularly impactful for tasks with a large result state space, such as those involving numeric answers. Our work is also significant because both quantitative analyses and manual evaluations reveal the observed gains stem from the models' ability to refine an initial reasoning chain by generating a second, improved chain within the same inference step, demonstrating previously elusive self-improvement. Our code and data are publicly available.

CoT-based Synthesizer: Enhancing LLM Performance through Answer Synthesis

Bohan Zhang, Xiaokang Zhang, Jing Zhang, Jifan Yu, Sijia Luo, Jie Tang

Current inference scaling methods, such as Self-consistency and Best-of-N, have proven effective in improving the accuracy of LLMs on complex reasoning tasks. However, these methods rely heavily on the quality of candidate responses and are unable to produce correct answers when all candidates are incorrect. In this paper, we propose a novel inference scaling strategy, CoT-based Synthesizer, which leverages CoT reasoning to synthesize superior answers by analyzing complementary information from multiple candidate responses, even when all candidates are flawed. To support a lightweight and cost-effective implementation, we introduce an automated data generation pipeline that creates diverse training data. This enables smaller LLMs trained on this data to improve the inference accuracy of larger models, including API-based LLMs. Experimental results across four benchmark datasets with seven policy models demonstrate that our method significantly enhances performance, with gains of 11.8% for Llama3-8B and 10.3% for GPT-4o on the MATH dataset. The corresponding training data and code are publicly available on the [repository](https://github.com/RUCKBReasoning/CoT-based-Synthesizer).

Can Graph Descriptive Order Affect Solving Graph Problems with LLMs?

Yuyao Ge, Shenghua Liu, Baolong Bi, Yiwei Wang, Lingrui Mei, Wenjie Feng, Lizhe Chen, Xueqi Cheng

Large language models (LLMs) have achieved significant success in reasoning tasks, including mathematical reasoning and logical deduction. Among these reasoning tasks, graph problems stand out due to their complexity and unique structural characteristics, attracting considerable attention from researchers. Previous studies have explored LLMs' graph reasoning abilities through various techniques, such as different encoding methods for graph structures and the use of carefully designed prompts. However, a critical factor has been mostly overlooked: the prompt sequential order in which graph descriptions are presented to the models. In this study, we present the first comprehensive analysis of how the order of graph descriptions impacts LLM performance. Specifically, we comprehensively evaluate four graph description orders across six graph problems using six mainstream LLMs. The results reveal that: (1) ordered graph descriptions significantly improve LLMs' comprehension of graph structures; (2) the robustness of LLMs to graph description order varies across different tasks; and (3) the impact of graph order on performance is closely related to the inherent characteristics of tasks. This study provides a critical advancement in the application of LLMs for solving graph-related problems, paving the way for future research to optimize model performance through strategic graph description ordering.

VQAGuider: Guiding Multimodal Large Language Models to Answer Complex Video Questions

Yuyan Chen, Jiyuan Jia, Jiaxin Lu, Siyue Li, Yu Guan, Ming Yang, Qingpei Guo

Complex video question-answering (VQA) requires in-depth understanding of video contents including object and action recognition as well as video classification and summarization, which exhibits great potential in emerging applications in education and entertainment, etc. Multimodal large language models (MLLMs) may accomplish this task by grasping the intention of a question and decomposing it to a series of visual recognition sub-tasks to find out the answer with the help of an agent. To tackle this task, we first collect a new dedicated Complex VQA dataset named CVQA and then propose VQAGuider, an innovative framework planning a few atomic visual recognition tools by video-related API matching. VQAGuider facilitates a deep engagement with video content and precise responses to complex video-related questions by MLLMs, which is beyond aligning visual and language features for simple VQA tasks. Our experiments demonstrate VQAGuider is capable of navigating the complex VQA tasks by MLLMs and improves the accuracy by 29.6\% and 17.2\% on CVQA and the existing VQA datasets, respectively, highlighting its potential in advancing MLLMs's capabilities in video understanding.

Can Large Language Models Accurately Generate Answer Keys for Health-related Questions?

Davis Bartels, Deepak Gupta, Dina Demner-Fushman

The evaluation of text generated by LLMs remains a challenge for question answering, retrieval augmented generation (RAG), summarization, and many other natural language processing tasks. Evaluating the factuality of LLM generated responses is particularly important in medical question answering, where the stakes are high. One method of evaluating the factuality of text is through the use of information nuggets (answer keys). Nuggets are text representing atomic facts that may be used by an assessor to make a binary decision as to whether the fact represented by said nugget is contained in an answer. Although manual nugget extraction is expensive and time-consuming, recent RAG shared task evaluations have explored automating the nuggetization of text with LLMs. In this work, we explore several approaches to nugget generation for medical question answering and evaluate their alignment with expert human nugget generation. We find providing an example and extracting nuggets from an answer to be the best approach to nuggetization. While, overall, we found the capabilities of LLMs to distill atomic facts limited, Llama 3.3 performed the best out of the models we tested.

CypherBench: Towards Precise Retrieval over Full-scale Modern Knowledge Graphs in the LLM Era

Yanlin Feng, Simone Papicchio, Sajjadur Rahman

Retrieval from graph data is crucial for augmenting large language models (LLM) with both open-domain knowledge and private enterprise data, and it is also a key component in the recent GraphRAG system \cite{edge2024local}. Despite decades of research on knowledge graphs and knowledge base question answering, leading LLM frameworks (\eg Langchain and LlamaIndex) have only minimal support for retrieval from modern encyclopedic knowledge graphs like Wikidata. In this paper, we analyze the root cause and suggest that modern RDF knowledge graphs (\eg Wikidata, Freebase) are less efficient for LLMs due to overly large schemas that far exceed the typical LLM context window, use of resource identifiers, overlapping and ambiguous relation types and lack of normalization. As a solution, we propose \textit{property graph views} on top of the underlying RDF graph that can be efficiently queried by LLMs using \textit{Cypher}. We instantiated this idea on Wikidata and introduced {CypherBench}, the first benchmark with 11 large-scale, multi-domain property graphs with 7.8 million entities and over 10,000 questions. To achieve this, we tackled several key challenges, including developing an RDF-to-property graph conversion engine, creating a systematic pipeline for text-to-Cypher task generation, and designing new evaluation metrics.

ARise: Towards Knowledge-Augmented Reasoning via Risk-Adaptive Search

Yize Zhang, Tianshu Wang, Sirui Chen, Kun Wang, Xingyu Zeng, Hongyu Lin, Xianpei Han, Le Sun, Chaochao Lu

Large language models (LLMs) have demonstrated impressive capabilities and are receiving increasing attention to enhance their reasoning through scaling test-time compute. However, their application in open-ended, knowledge-intensive, complex reasoning scenarios is still limited. Reasoning-oriented methods struggle to generalize to open-ended scenarios due to implicit assumptions of complete world knowledge. Meanwhile, knowledge-augmented reasoning (KAR) methods fails to address two core challenges: 1) error propagation, where errors in early steps cascade through the chain, and 2) verification bottleneck, where the explore–exploit trade-off arises in multi-branch decision processes. To overcome these limitations, we introduce ARise, a novel framework that integrates risk assessment of intermediate reasoning states with dynamic retrieval-augmented generation (RAG) within a Monte Carlo tree search paradigm. This approach enables effective construction and optimization of reasoning plans across multiple maintained hypothesis branches. Experimental results show that ARise significantly outperforms the state-of-the-art KAR methods by up to 23.10%, and the latest RAG-equipped large reasoning models by up to 25.37%. Our project page is at https://opencausalab.github.io/ARise.

Safe: Enhancing Mathematical Reasoning in Large Language Models via Retrospective Step-aware Formal Verification

Chengwu Liu, Ye Yuan, Yichun Yin, Yan Xu, Xin Xu, Zaoyu Chen, Lifeng Shang, Qun Liu, Ming Zhang

Chain-of-Thought (CoT) prompting has become the de facto method to elicit reasoning capabilities from large language models (LLMs). However, to mitigate hallucinations in CoT that are notoriously difficult to detect, current methods such as process reward models (PRMs) or self-consistency operate as opaque boxes and do not provide checkable evidence for their judgments, possibly limiting their effectiveness. To address this issue, we draw inspiration from the idea that "the gold standard for supporting a mathematical claim is to provide a proof". We propose a retrospective, step-aware formal verification framework Safe. Rather than assigning arbitrary scores, we strive to articulate mathematical claims in formal mathematical language Lean 4 at each reasoning step and provide formal proofs to identify hallucinations. We evaluate our framework Safe across multiple language models and various mathematical datasets, demonstrating a significant performance improvement while offering interpretable and verifiable evidence. We also propose FormalStep as a benchmark for step correctness theorem proving with 30,809 formal statements. To the best of our knowledge, our work represents the first endeavor to utilize formal mathematical language Lean 4 for verifying content generated by LLMs, aligning with the reason why formal mathematical languages were created in the first place: to provide a robust foundation for hallucination-prone human-written proofs.

Extending Complex Logical Queries on Uncertain Knowledge Graphs

Weizhi Fei, Zihao Wang, Hang Yin, Yang Duan, Yangqiu Song

The study of machine learning-based logical query-answering enables reasoning with large-scale and incomplete knowledge graphs. This paper further advances this line of research by considering the uncertainty in the knowledge. The uncertain nature of knowledge is widely observed in the real world, but does not align seamlessly with the first-order logic underpinning existing studies. To bridge this gap, we study the setting of soft queries on uncertain knowledge, which is motivated by the establishment of soft constraint programming. We further propose an ML-based approach with both forward inference and backward calibration to answer soft queries on large-scale, incomplete, and uncertain knowledge graphs. Theoretical discussions reveal that our method ensures there are no catastrophic cascading errors in our forward inference algorithm while maintaining the same complexity as state-of-the-art inference algorithms for first-order queries. Empirical results justify the superior performance of our approach against previous ML-based methods with number embedding extensions.

Fast or Slow? Integrating Fast Intuition and Deliberate Thinking for Enhancing Visual Question Answering

Songtao Jiang, Chenyi Zhou, Yan Zhang, Yeying Jin, Zuozhu Liu

Multimodal large language models (MLLMs) still struggle with complex reasoning tasks in Visual Question Answering (VQA). While current methods have advanced by incorporating visual prompts, our study uncovers critical limitations: these approaches indiscriminately annotate all detected objects for every visual question, generating excessive visual markers that degrade task performance. This issue stems primarily from a lack of focus on key visual elements, raising two important questions: Are all objects equally important, and do all questions require visual prompts? Motivated by Dual Process Theory, which distinguishes between instinctive and deliberate cognitive modes in human reasoning, we propose FOCUS, a plug-and-play approach that dynamically adapts to the complexity of questions, combining fast intuitive judgments with deliberate analytical reasoning to enhance the vision-language reasoning capability of the MLLM. For straightforward questions, FOCUS supports efficient zero-shot reasoning. For more complex tasks, it employs the conceptualizing before observation strategy to highlight critical elements. Extensive experiments on four benchmarks—ScienceQA, TextQA, VizWiz, and MME—demonstrate that FOCUS consistently improves the performance of both open-source and black-box MLLMs, achieving significant gains across all datasets. Ablation studies further validate the importance of combining diverse cognitive strategies with refined visual information for superior performance. Code will be released.

Ontology-Guided Reverse Thinking Makes Large Language Models Stronger on Knowledge Graph Question Answering

Runxuan Liu, luobei, Jiaqi Li, Baoxin Wang, Ming Liu, Dayong Wu, Shijin Wang, Bing Qin

Large language models (LLMs) have shown remarkable capabilities in natural language processing. However, in knowledge graph question answering tasks (KGQA), there remains the issue of answering questions that require multi-hop reasoning. Existing methods rely on entity vector matching, but the purpose of the question is abstract and difficult to match with specific entities. As a result, it is difficult to establish reasoning paths to the purpose, which leads to information loss and redundancy. To address this issue, inspired by human reverse thinking, we propose Ontology-Guided Reverse Thinking (ORT), a novel framework that constructs reasoning paths from purposes back to conditions. ORT operates in three key phases: (1) using LLM to extract purpose labels and condition labels, (2) constructing label reasoning paths based on the KG ontology, and (3) using the label reasoning paths to guide knowledge retrieval. Experiments on the WebQSP and CWQ datasets show that ORT achieves state-of-the-art performance and significantly enhances the capability of LLMs for KGQA.

Digest the Knowledge: Large Language Models empowered Message Passing for Knowledge Graph Question Answering

Junhong Wan, Tao Yu, Kunyu Jiang, Yao Fu, Weihao Jiang, Jiang Zhu

Despite their success, large language models (LLMs) suffer from notorious hallucination issue. By introducing external knowledge stored in knowledge graphs (KGs), existing methods use paths as the medium to represent the graph information that send into LLMs. However, paths only contain limited graph structure information and are unorganized with redundant sequentially appeared keywords, which are difficult for LLMs to digest. We aim to find a suitable medium that captures the essence of structure knowledge in KGs. Inspired by the Neural Message Passing in Graph Neural Networks, we propose Language Message Passing (LMP) that first learns a concise facts graph by iteratively aggregates neighbor entities and transforms them into semantic facts, and then we performs Topological Readout that encodes the graph structure information into multi-level lists of texts to augment LLMs. Our method serves as a brand-new innovative framework that brings a new perspective into KG-enhanced LLMs, and also offers human-level semantic explainability with significant performance improvements over existing methods on all 5 knowledge graph question answering datasets. Code is available at https://github.com/wanjunhong0/LMP.

iQUEST: An Iterative Question-Guided Framework for Knowledge Base Question Answering

Shuai Wang, Yinan Yu

While Large Language Models (LLMs) excel at many natural language processing tasks, they often suffer from factual inaccuracies in knowledge-intensive scenarios. Integrating external knowledge resources, particularly knowledge graphs (KGs), provides a transparent and updatable foundation for more reliable reasoning. Knowledge Base Question Answering (KBQA), which queries and reasons over KGs, is central to this effort, especially for complex, multi-hop queries. However, multi-hop reasoning poses two key challenges: (1) maintaining coherent reasoning paths, and (2) avoiding prematurely discarding critical multi-hop connections. To address these issues, we introduce iQUEST, a question-guided KBQA framework that iteratively decomposes complex queries into simpler sub-questions, ensuring a structured and focused reasoning trajectory. Additionally, we integrate a Graph Neural Network (GNN) to look ahead and incorporate 2-hop neighbor information at each reasoning step. This dual approach strengthens the reasoning process, enabling the model to explore viable paths more effectively. Detailed experiments demonstrate the consistent improvement delivered by iQUEST across four benchmark datasets and four LLMs.

ChronoSense: Exploring Temporal Understanding in Large Language Models with Time Intervals of Events

Duygu Sezen Islakoglu, Jan-Christoph Kalo

Large Language Models (LLMs) still face significant challenges in reasoning and arithmetic. Although temporal reasoning has raised increasing research attention, comprehensive testing of Allen's interval relations (e.g., before, after, during) —a fundamental framework for temporal relationships— remains underexplored. To fill this gap, we present ChronoSense, a new benchmark for evaluating LLMs' temporal understanding. It includes 16 tasks, identifying the Allen relation between two temporal events and temporal arithmetic. We assess the performance of seven recent LLMs. The results indicate that models handle Allen relations, even symmetrical ones, quite differently. Moreover, the findings suggest that the models may rely on memorization to answer time-related questions. Overall, the models' low performance highlights the need for improved temporal understanding in LLMs. Our dataset and the source code are available at https://github.com/duyguislakoglu/chronosense.

SConU: Selective Conformal Uncertainty in Large Language Models

Zhiyuan Wang, Qingni Wang, Yue Zhang, Tianlong Chen, Xiaofeng Zhu, Xiaoshuang Shi, Kaidi Xu

As large language models are increasingly utilized in real-world applications, guarantees of task-specific metrics are essential for their reliable deployment. Previous studies have introduced various criteria of conformal uncertainty grounded in split conformal prediction, which offer user-specified correctness coverage. However, existing frameworks often fail to identify uncertainty data outliers that violate the exchangeability assumption, leading to unbounded miscoverage rates and unactionable prediction sets. In this paper, we propose a novel approach termed Selective Conformal Uncertainty (SConU), which, for the first time, implements significance tests, by developing two conformal p-values that are instrumental in determining whether a given sample deviates from the uncertainty distribution of the calibration set at a specific manageable risk level. Our approach not only facilitates rigorous management of miscoverage rates across both single-domain and interdisciplinary contexts, but also enhances the efficiency of predictions. Furthermore, we comprehensively analyze the components of the conformal procedures, aiming to approximate conditional coverage, particularly in high-stakes question-answering tasks.

Beyond Frameworks: Unpacking Collaboration Strategies in Multi-Agent Systems

Haochun Wang, Sendong Zhao, Jingbo Wang, Zewen Qiang, Bing Qin, Ting Liu

Multi-agent collaboration has emerged as a pivotal paradigm for addressing complex, distributed tasks in large language model (LLM)-driven applications. While prior research has focused on high-level architectural frameworks, the granular mechanisms governing agents—critical to performance and scalability—remain underexplored. This study systematically investigates four dimensions of collaboration strategies: (1) agent governance, (2) participation control, (3) interaction dynamics, and (4) dialogue history management. Through rigorous experimentation under two context-dependent scenarios—Distributed Evidence Integration (DEI) and Structured Evidence Synthesis (SES)—we quantify the impact of these strategies on both task accuracy and computational efficiency. Our findings reveal that centralized governance, instructor-led participation, ordered interaction patterns, and instructor-curated context summarization collectively optimize the trade-off between decision quality and resource utilization with the support of the proposed Token-Accuracy Ratio (TAR). This work establishes a foundation for designing adaptive, scalable multi-agent systems, shifting the focus from structural novelty to strategic interaction mechanics.

Exclusion of Thought: Mitigating Cognitive Load in Large Language Models for Enhanced Reasoning in Multiple-Choice Tasks

Qihang Fu, Yongbin Qin, Ruizhang Huang, Yanping Chen, Yulin Zhou, Lintao Long

Multiple-choice questions (MCQs) are a widely used and vital assessment format for evaluating large language models (LLMs). This study reveals that LLMs are susceptible to "cognitive load" caused by distractor options in MCQs, leading to excessive attention to distractors and consequent vacillation between correct and incorrect options. To mitigate this cognitive burden, we introduce a novel reasoning prompt strategy, called EoT, which effectively reduces cognitive load by steering the model’s attention away from erroneous options. This enables the model to focus more effectively on reasonable answers. Additionally, by documenting the elimination process, EoT enhances the transparency and interpretability of the model's reasoning. Experimental results demonstrate that EoT, as a plug-and-play approach, significantly reduces cognitive load and improves performance, showcasing its potential to enhance both the accuracy and interpretability of LLMs.

Finding Needles in Images: Can Multi-modal LLMs Locate Fine Details?

Parth Thakkar, Ankush Agarwal, Prasad Kasu, Pulkit Bansal, Chaitanya Devaguptapu

While Multi-modal Large Language Models (MLLMs) have shown impressive capabilities in document understanding tasks, their ability to locate and reason about fine-grained details within complex documents remains understudied. Consider searching a restaurant menu for a specific nutritional detail or identifying a disclaimer in a lengthy newspaper article — tasks that demand careful attention to small but significant details within a broader narrative, akin to Finding Needles in Images (NiM). To address this gap, we introduce NiM-Benchmark, a carefully curated benchmark spanning diverse real-world documents including newspapers, menus, and lecture images, specifically designed to evaluate MLLMs' capability in these intricate tasks. Building on this, we further propose Spot-IT, a simple yet effective approach that enhances MLLMs capability through intelligent patch selection and Gaussian attention, motivated from how humans zoom and focus when searching documents. Our extensive experiments reveal both the capabilities and limitations of current MLLMs in handling fine-grained document understanding tasks, while demonstrating the effectiveness of our approach. Spot-IT achieves significant improvements over baseline methods, particularly in scenarios requiring precise detail extraction from complex layouts.

Towards Fully Exploiting LLM Internal States to Enhance Knowledge Boundary Perception

Shiyu Ni, Keping Bi, Jiafeng Guo, Lulu Yu, Baolong Bi, Xueqi Cheng

Large language models (LLMs) exhibit impressive performance across diverse tasks but often struggle to accurately gauge their knowledge boundaries, leading to confident yet incorrect responses. This paper explores leveraging LLMs' internal states to enhance their perception of knowledge boundaries from efficiency and risk perspectives. We investigate whether LLMs can estimate their confidence using internal states before response generation, potentially saving computational resources. Our experiments on datasets like Natural Questions, HotpotQA, and MMLU reveal that LLMs demonstrate significant pre-generation perception, which is further refined post-generation, with perception gaps remaining stable across varying conditions. To mitigate risks in critical domains, we introduce Consistency-based Confidence Calibration (C^3), which assesses confidence consistency through question reformulation. C^3 significantly improves LLMs' ability to recognize their knowledge gaps, enhancing the unknown perception rate by 5.6\% on NQ and 4.9\% on HotpotQA. Our findings suggest that pre-generation confidence estimation can optimize efficiency, while C^3 effectively controls output risks, advancing the reliability of LLMs in practical applications.

Chain-of-Reasoning: Towards Unified Mathematical Reasoning in Large Language Models via a Multi-Paradigm Perspective

Yiyao Yu, Yuxiang Zhang, Dongdong Zhang, Xiao Liang, Hengyuan Zhang, Xingxing Zhang, MAHMOUD KHADEMI, Hany Hassan Awadalla, Junjie Wang, Yujiu Yang, Furu Wei

Large Language Models (LLMs) have made notable progress in mathematical reasoning, yet they often rely on single-paradigm reasoning that limits their effectiveness across diverse tasks. In this paper, we introduce Chain-of-Reasoning (CoR), a novel unified framework that integrates multiple reasoning paradigms --- Natural Language Reasoning (NLR), Algorithmic Reasoning (AR), and Symbolic Reasoning (SR) --- to enable synergistic collaboration. CoR generates multiple potential answers using different reasoning paradigms and synthesizes them into a coherent final solution. We propose a Progressive Paradigm Training (PPT) strategy that allows models to progressively master these paradigms, culminating in the development of at CoR-Math-7B. Experimental results demonstrate that CoR-Math-7B significantly outperforms current SOTA models, achieving up to a 41.0% absolute improvement over GPT-4o in theorem proving tasks and a 15% improvement over RL-based methods on the MATH benchmark in arithmetic tasks. These results show the enhanced mathematical comprehensive ability of our model, enabling zero-shot generalization across tasks.The code is available at https://github.com/microsoft/CoR.

Answering Complex Geographic Questions by Adaptive Reasoning with Visual Context and External Commonsense Knowledge

Fan Li, Jianxing Yu, Jielong Tang, Wenqing Chen, Hanjiang Lai, Yanghui Rao, Jian Yin

This paper focuses on a new task of answering geographic reasoning questions based on the given image (called GeoVQA). Unlike traditional VQA tasks, GeoVQA asks for details about the image-related culture, landscape, etc. This requires not only the identification of the objects in the image, their properties and relations, but also the understanding of the geographic knowledge of the objects, such as location, transportation, landmark, cuisine, etc. This background knowledge does not explicitly appear in the image, nor is there an extra-textual description. Without this missing but necessary knowledge, it is difficult for existing matching-based methods to infer the correct answer. To tackle these challenges, we propose a new geographic reasoning framework for our task. We first analyze the image and describe its fine-grained content by text and keywords using a multi-modal retrieval augmented technique, so as to deduce an answer in a unified textual modality. Next, we retrieve the crucial geographic commonsense knowledge. To reduce the retrieval complexity, we design a dynamic method that can adaptively collect the relevant clues for each reasoning step. The step in the incorrect direction will be pruned according to some judgment criteria. The remaining steps can help us form a reasoning chain to derive a correct answer. Moreover, we create a large-scale dataset GVQA with 41,329 samples to conduct the evaluation. The results demonstrate the effectiveness of our approach.

Exposing the Achilles' Heel: Evaluating LLMs Ability to Handle Mistakes in Mathematical Reasoning

Joykirat Singh, Akshay Nambi, Vibhav Vineet

Large Language Models (LLMs) have significantly impacted the field of Math Word Problems (MWPs), transforming how these problems are approached and solved, particularly in educational contexts. However, existing evaluations often focus on final accuracy, neglecting the critical aspect of reasoning capabilities. This work addresses that gap by evaluating LLMs' abilities to detect and correct reasoning mistakes. We present a novel dataset, MWP-MISTAKE, containing MWPs with both correct and incorrect reasoning steps generated through rule-based methods and smaller language models. Our comprehensive benchmarking of state-of-the-art models such as GPT-4o and GPT4 uncovers important insights into their strengths and limitations. While GPT-4o excels in mistake detection and rectification, gaps remain, particularly in handling complex datasets and novel problems. Additionally, we identify concerns with data contamination and memorization, which affect LLM reliability in real-world applications. While OpenAI' O1 model demonstrates 90% accuracy in reasoning and final answers on complex tasks, it remains weak in mistake detection. Our findings highlight the need for improved reasoning evaluations and suggest ways to enhance LLM generalization and robustness in math problem-solving.

GRAT: Guiding Retrieval-Augmented Reasoning through Process Rewards Tree Search

Xianshu Peng, Wei Wei

Enhancing large models for complex multi-hop question-answering has become a research focus in the Retrieval-augmented generation (RAG) area. Many existing approaches aim to mimic human thought processes by enabling large models to perform retrieval-augmented generation step by step. However, these methods can only perform single chain reasoning, which lacks the ability for multi-path exploration, strategic look-ahead, stepwise evaluation, and global selection. In addition, to effectively decompose complex problems, these methods can only rely on labor-intensive intermediate annotations for supervised fine-tuning. To address these issues, we propose GRAT, an algorithm guided by Monte Carlo Tree Search (MCTS) and process rewards. GRAT not only enables self-evaluation and self-correction but also assigns fine-grained rewards to each intermediate step in the search path. These fine-grained annotations can be used for model self-training, which enables GRAT to continuously self-update its problem analysis and reasoning capabilities. We conducted experiments on four multihop QA datasets: HotPotQA, 2WikiMultiHopQA, MuSiQue, and Bamboogle, demonstrating that GRAT outperforms various RAG-based methods. Additionally, incorporating self-training significantly enhances GRAT's reasoning performance.

Exploiting Contextual Knowledge in LLMs through V-usable Information based Layer Enhancement

Xiaowei Yuan, Zhao Yang, Ziyang Huang, Yequan Wang, Siqi Fan, Yiming Ju, Jun Zhao, Kang Liu

Large Language Models (LLMs) have demonstrated remarkable capabilities in various tasks, yet they often struggle with context-faithfulness generations that properly reflect contextual knowledge. While existing approaches focus on enhancing the decoding strategies, they ignore the fundamental mechanism of how contextual information is processed within LLMs' internal states. As a result, LLMs remain limited in their ability to fully leverage contextual knowledge. In this paper, we propose Context-aware Layer Enhancement (CaLE), a novel intervention method that enhances the utilization of contextual knowledge within LLMs' internal representations. By employing \mathcal{V}-usable information analysis, CaLE strategically amplifies the growth of contextual information at an optimal layer, thereby enriching representations in the final layer. Our experiments demonstrate that CaLE effectively improves context-faithful generation in Question-Answering tasks, particularly in scenarios involving unknown or conflicting contextual knowledge.

Dynamic Chunking and Selection for Reading Comprehension of Ultra-Long Context in Large Language Models

Boheng Sheng, Jiacheng Yao, Meicong Zhang, Guoxiu He

Large language models (LLMs) often struggle to accurately read and comprehend extremely long texts. Current methods for improvement typically rely on splitting long contexts into fixed-length chunks. However, fixed truncation risks separating semantically relevant content, leading to ambiguity and compromising accurate understanding. To overcome this limitation, we propose a straightforward approach for dynamically separating and selecting chunks of long context, facilitating a more streamlined input for LLMs. In particular, we compute semantic similarities between adjacent sentences, using lower similarities to adaptively divide long contexts into variable-length chunks. We further train a question-aware classifier to select sensitive chunks that are critical for answering specific questions. Experimental results on both single-hop and multi-hop question-answering benchmarks show that the proposed approach consistently outperforms strong baselines. Notably, it maintains robustness across a wide range of input lengths, handling sequences of up to 256k tokens. Our datasets and code are available at the following link: https://github.com/ECNU-Text-Computing/DCS

HELIOS: Harmonizing Early Fusion, Late Fusion, and LLM Reasoning for Multi-Granular Table-Text Retrieval

Sungho Park, Joohyung Yun, Jongwuk Lee, Wook-Shin Han

Table-text retrieval aims to retrieve relevant tables and text to support open-domain question answering. Existing studies use either early or late fusion, but face limitations. Early fusion pre-aligns a table row with its associated passages, forming ``stars,'' which often include irrelevant contexts and miss query-dependent relationships. Late fusion retrieves individual nodes, dynamically aligning them, but it risks missing relevant contexts. Both approaches also struggle with advanced reasoning tasks, such as column-wise aggregation and multi-hop reasoning. To address these issues, we propose HELIOS, which combines the strengths of both approaches. First, the edge-based bipartite subgraph retrieval identifies finer-grained edges between table segments and passages, effectively avoiding the inclusion of irrelevant contexts. Then, the query-relevant node expansion identifies the most promising nodes, dynamically retrieving relevant edges to grow the bipartite subgraph, minimizing the risk of missing important contexts. Lastly, the star-based LLM refinement performs logical inference at the star graph level rather than the bipartite subgraph, supporting advanced reasoning tasks. Experimental results show that HELIOS outperforms state-of-the-art models with a significant improvement up to 42.6% and 39.9% in recall and nDCG, respectively, on the OTT-QA benchmark.

CoRe-MMRAG: Cross-Source Knowledge Reconciliation for Multimodal RAG

yang tian, Fan Liu, Jingyuan Zhang, V. W., Yupeng Hu, Liqiang Nie

Multimodal Retrieval-Augmented Generation (MMRAG) has been introduced to enhance Multimodal Large Language Models by incorporating externally retrieved multimodal knowledge, but it introduces two challenges: Parametric-Retrieved Knowledge Inconsistency (PRKI), where discrepancies between parametric and retrieved knowledge create uncertainty in determining reliability, and Visual-Textual Knowledge Inconsistency (VTKI), where misalignment between visual and textual sources disrupts entity representation. To address these challenges, we propose \textbf{C}r\textbf{o}ss-source knowledge \textbf{Re}conciliation for \textbf{M}ulti\textbf{M}odal \textbf{RAG} (CoRe-MMRAG), a novel end-to-end framework that effectively reconciles inconsistencies across knowledge sources. CoRe-MMRAG follows a four-stage pipeline: it first generates an internal response from parametric knowledge, then selects the most relevant multimodal evidence via joint similarity assessment, generates an external response, and finally integrates both to produce a reliable answer. Additionally, a specialized training paradigm enhances knowledge source discrimination, multimodal integration, and unified answer generation. Experiments on KB-VQA benchmarks show that CoRe-MMRAG achieves substantial improvements over baseline methods, achieving 5.6\% and 9.3\% performance gains on InfoSeek and Encyclopedic-VQA, respectively. We release code and data at \href{https://github.com/TyangJN/CoRe-MMRAG}{https://github.com/TyangJN/CoRe-MMRAG}.

Session 12: Question Answering (Wednesday 30 July, 11:00-12:30, Hall 5X)

My Climate CoPilot: A Question Answering System for Climate Adaptation in Agriculture

Vincent Nguyen, Willow Hallgren, Ashley Harkin, Mahesh Prakash, Sarvnaz Karimi

Accurately answering climate science questions requires scientific literature and climate data. Interpreting climate literature and data, however, presents inherent challenges, such as interpreting uncertainties, determining relevant climate factors and drivers, and dealing with the sheer volume of data. My Climate CoPilot is a platform that assists climate experts, such as farmer advisors, mitigate and adapt to projected climate impacts by answering questions grounded in evidence. It emphasises transparency, user privacy, and low-resource use, and provides automatic evaluation. Fifty domain experts have carefully evaluated every aspect of My Climate CoPilot and based on their interactions and feedback, the system continues to evolve. To date, we have received 2180 question-answer pairs with an 82% positive rating from the experts.

CiteLab: Developing and Diagnosing LLM Citation Generation Workflows via the Human-LLM Interaction

Jiajun Shen, Tong Zhou, Yubo Chen, Kang Liu, Jun Zhao

The emerging paradigm of enabling Large Language Models (LLMs) to generate citations in Question-Answering (QA) tasks is lacking in a unified framework to standardize and fairly compare different citation generation methods, leading to difficulties in reproduction and innovation. Therefore, we introduce Citeflow, an open-source and modular framework fostering reproduction and the implementation of new designs. Citeflow is highly extensible, allowing users to utilize four main modules and 14 components to construct a pipeline, evaluate an existing method, and understand the attributing LLM-generated contents. The framework is also paired with a visual interface, Citefix, facilitating case study and modification of existing citation generation methods. Users can use this interface to conduct LLM-powered case studies according to different scenarios. Citeflow and Citefix are highly integrated into the toolkit CiteLab, and we use an authentic process of multiple rounds of improvement through the Human-LLM interaction interface to demonstrate the efficiency of our toolkit on implementing and modifying citation generation pipelines. Citelab is released at https://github.com/SjJ1017/Citelab

Ai2 Scholar QA: Organized Literature Synthesis with Attribution

Amanpreet Singh, Joseph Chee Chang, Dany Haddad, Aakanksha Naik, Jena D. Hwang, Rodney Kinney, Daniel S Weld, Doug Downey, Sergey Feldman

Retrieval-augmented generation is increasingly effective in answering scientific questions from literature, but many state-of-the-art systems are expensive and closed-source. We introduce Ai2 Scholar QA, a free online scientific question answering application. To facilitate research, we make our entire pipeline public: as a customizable open-source Python package and interactive web app, along with paper indexes accessible through public APIs and downloadable datasets. We describe our system in detail and present experiments analyzing its key design decisions. In an evaluation on a recent scientific QA benchmark, we find that Ai2 Scholar QA outperforms competing systems.

Session 12: Resources and Evaluation (Wednesday 30 July, 11:00-12:30, Hall 4/5)

FactLens: Benchmarking Fine-Grained Fact Verification

Kushan Mitra, Dan Zhang, Sajjadur Rahman, Estevam Hruschka

Large Language Models (LLMs) have shown impressive capability in language generation and understanding, but their tendency to hallucinate and produce factually incorrect information remains a key limitation. To verify LLM-generated contents and claims from other sources, traditional verification approaches often rely on holistic models that assign a single factuality label to complex claims, potentially obscuring nuanced errors. In this paper, we advocate for a shift toward fine-grained verification, where complex claims are broken down into smaller sub-claims for individual verification, allowing for more precise identification of inaccuracies, improved transparency, and reduced ambiguity in evidence retrieval. However, generating sub-claims poses challenges, such as maintaining context and ensuring semantic equivalence with respect to the original claim. We introduce **FactLens**, a benchmark for evaluating fine-grained fact verification, with metrics and automated evaluators of sub-claim quality. The benchmark data is manually curated to ensure high-quality ground truth. Our results show alignment between automated FactLens evaluators and human judgments, and we discuss the impact of sub-claim characteristics on the overall verification performance.

MultiChallenge: A Realistic Multi-Turn Conversation Evaluation Benchmark Challenging to Frontier LLMs

Kaustubh Deshpande, Ved Sirdeshmukh, Johannes Baptist Mols, Lifeng Jin, Ed-Yeremai Hernandez-Cardona, Dean Lee, Jeremy Kritz, Willow E. Primack, Summer Yue, Chen Xing

We present MultiChallenge, a pioneering benchmark evaluating large language models (LLMs) on conducting multi-turn conversations with human users, a crucial yet underexamined capability for their applications. MultiChallenge identifies four categories of challenges in multi-turn conversations that are not only common and realistic among current human-LLM interactions, but are also challenging to all current frontier LLMs. All 4 challenges require accurate instruction-following, context allocation, and in-context reasoning at the same time. We also develop LLM as judge with instance-level rubrics to facilitate an automatic evaluation method with fair agreement with experienced human raters. Despite achieving near perfect scores on existing multi-turn evaluation benchmarks, all frontier models have less than 50\% accuracy on MultiChallenge, with the top-performing Claude 3.5 Sonnet (October 2024) achieving just a 41.4\% average accuracy.

MVTamperBench: Evaluating Robustness of Vision-Language Models

Amit Agarwal, Srikant Panda, Angeline Charles, Hitesh Laxmichand Patel, Bhargava Kumar, Priyaranjan Pattnayak, Taki Hasan Rafi, Tejaswini Kumar, Hansa Meghwani, Karan Gupta, Dong-Kyu Chae

Multimodal Large Language Models (MLLMs), are recent advancement of Vision-Language Models (VLMs) that have driven major advances in video understanding. However, their vulnerability to adversarial tampering and manipulations remains underexplored. To address this gap, we introduce \textbf{MVTamperBench}, a benchmark that systematically evaluates MLLM robustness against five prevalent tampering techniques: rotation, masking, substitution, repetition, and dropping; based on real-world visual tampering scenarios such as surveillance interference, social media content edits, and misinformation injection. MVTamperBench comprises 3.4K original videos, expanded into over 17K tampered clips covering 19 distinct video manipulation tasks. This benchmark challenges models to detect manipulations in spatial and temporal coherence. We evaluate 45 recent MLLMs from 15+ model families. We reveal substantial variability in resilience across tampering types and show that larger parameter counts do not necessarily guarantee robustness. MVTamperBench sets a new benchmark for developing tamper-resilient MLLM in safety-critical applications, including detecting clickbait, preventing harmful content distribution, and enforcing policies on media platforms. We release all code, data, and benchmark to foster open research in trustworthy video understanding.

A Multi-Labeled Dataset for Indonesian Discourse: Examining Toxicity, Polarization, and Demographics Information

Lucky Susanto, Musa Izzanardi Wijanarko, Prasetia Anugrah Pratama, Zilu Tang, Fariz Akyas, Traci Hong, Ika Karlina Idris, Alham Fikri Aji, Derry Tanti Wijaya

Online discourse is increasingly trapped in a vicious cycle where polarizing language fuels toxicity and vice versa. Identity, one of the most divisive issues in modern politics, often increases polarization. Yet, prior NLP research has mostly treated toxicity and polarization as separate problems. In Indonesia, the world's third-largest democracy, this dynamic threatens democratic discourse, particularly in online spaces. We argue that polarization and toxicity must be studied in relation to each other. To this end, we present a novel multi-label Indonesian dataset annotated for toxicity, polarization, and annotator demographic information. Benchmarking with BERT-base models and large language models (LLMs) reveals that polarization cues improve toxicity classification and vice versa. Including demographic context further enhances polarization classification performance.

Disentangling Logic: The Role of Context in Large Language Model Reasoning Capabilities

Wenyue Hua, Kaijie Zhu, Lingyao Li, Lizhou Fan, Mingyu Jin, Shuhang Lin, Haochen Xue, Zelong Li, Jindong Wang, Yongfeng Zhang

This study intends to systematically disentangle pure logic reasoning and text understanding by investigating the contrast across abstract and contextualized logical problems from a comprehensive set of domains. We explore whether LLMs demonstrate genuine reasoning capabilities across various domains when the underlying logical structure remains constant. We focus on two main questions (1) Can abstract logical problems alone accurately benchmark LLMs' reasoning ability in real-world scenarios, disentangled from contextual support in practical settings? (2) Does fine-tuning LLMs on abstract logic problems generalize to contextualized logic problems and vice versa? To investigate these questions, we focus on standard propositional logic, specifically propositional deductive and abductive logic reasoning. We construct datasets for both reasoning types with four difficulty levels across 12 distinct domains based on the Wikipedia categorization in addition to those with purely abstract variables. Our experiments aim to provide insights into disentangling context in logical reasoning, the genuine reasoning capabilities of LLMs, and their generalization potential. Coda and data are available at .

DocMEdit: Towards Document-Level Model Editing

Li Zeng, Zeming Liu, Chong Feng, Heyan Huang, Yuhang Guo

Model editing aims to correct errors and outdated knowledge in the Large language models (LLMs) with minimal cost. Prior research has proposed a variety of datasets to assess the effectiveness of these model editing methods. However, most existing datasets only require models to output short phrases or sentences, which overlooks the widespread existence of document level tasks in the real world, raising doubts about their practical usability. Aimed at addressing this limitation and promoting the application of model editing in real-world scenarios, we propose the task of document-level model editing. To tackle such challenges and enhance model capabilities in practical settings, we introduce DocMEdit, a dataset focused on document-level model editing, characterized by document-level inputs and outputs, extrapolative, and multiple facts within a single edit. We propose a series of evaluation metrics and experiments. The results show that the difficulties in document-level model editing pose challenges for existing model editing methods.

V²R-Bench: Holistically Evaluating LVLM Robustness to Fundamental Visual Variations

Zhiyuan Fan, Yumeng Wang, Sandeep Polisetty, Yi R. Fung

Large Vision Language Models (LVLMs) have shown impressive performance on various vision-language tasks. However, while objects in natural scenes inevitably exhibit visual variations in position, scale, orientation, and context due to changes in viewpoint and environment, the robustness of LVLMs to these fundamental visual variations remains largely unexplored. To address this gap, we introduce V²R-Bench, a comprehensive benchmark framework for evaluating Visual Variation Robustness of LVLMs, which encompasses automated evaluation dataset generation and principled metrics for thorough robustness assessment. Through extensive evaluation of 13 LVLMs, we reveal a surprising vulnerability to visual variations, affecting even advanced models that excel at complex vision-language tasks yet significantly underperform on simple tasks like object recognition. Interestingly, these models exhibit a distinct visual position bias that contradicts theories of effective receptive fields and demonstrate a human-like visual acuity threshold. To identify the source of these vulnerabilities, we propose a systematic framework for component-level analysis, featuring a novel visualization approach for aligned visual features. Results show that these vulnerabilities stem from error accumulation in the pipeline architecture and inadequate multimodal alignment. Complementary experiments with synthetic data further demonstrate that these limitations are fundamentally architectural challenges, underscoring the need for architectural innovations in future LVLM designs.

KE-MHISTO: Towards a Multilingual Historical Knowledge Extraction Benchmark for Addressing the Long-Tail Problem

Arianna Graciotti, Leonardo Piano, Nicolas Lazzari, Enrico Daga, Rocco Tripodi, Valentina Presutti, Livio Pompianu

Large Language Models (LLMs) face significant challenges when queried about long-tail knowledge, i.e., information that is rarely encountered during their training process. These difficulties arise due to the inherent sparsity of such data. Furthermore, LLMs often lack the ability to verify or ground their responses in authoritative sources, which can lead to plausible yet inaccurate outputs when addressing infrequent subject matter. Our work aims to investigate these phenomena by introducing KE-mHISTO, a multilingual benchmark for Entity Linking and Question Answering in the domain of historical music knowledge, available in both Italian and English. We demonstrate that KE-mHISTO provides significantly broader coverage of long-tail knowledge compared to existing alternatives. Moreover, it poses substantial challenges for state-of-the-art models. Our experiments reveal that smaller, multilingual models can achieve performance comparable to significantly larger counterparts, highlighting the potential of efficient, language-aware approaches for long-tail knowledge extraction. KE-mHISTO is available at: https://anonymous.4open.science/r/KE-MHISTO-6108.

CLEAR: Character Unlearning in Textual and Visual Modalities

Alexey Dontsov, Dmitrii Korzh, Alexey Zhavoronkin, Boris Mikheev, Denis Bobkov, Aibek Alanov, Oleg Rogov, Ivan Oseledets, Elena Tutubalina

Machine Unlearning (MU) is critical for removing private or hazardous information from deep learning models. While MU has advanced significantly in unimodal (text or vision) settings, multimodal unlearning (MMU) remains underexplored due to the lack of open benchmarks for evaluating cross-modal data removal. To address this gap, we introduce CLEAR, the first open-source benchmark designed specifically for MMU. CLEAR contains 200 fictitious individuals and 3,700 images linked with corresponding question-answer pairs, enabling a thorough evaluation across modalities. We conduct a comprehensive analysis of 11 MU methods (e.g., SCRUB, gradient ascent, DPO) across four evaluation sets, demonstrating that jointly unlearning both modalities outperforms single-modality approaches.

Assessing the Reasoning Capabilities of LLMs in the context of Evidence-based Claim Verification

John Dougrez-Lewis, Mahmud Elahi Akhter, Federico Ruggeri, Sebastian Löbbers, Yulan He, Maria Liakata

Although LLMs have shown great performance on Mathematics and Coding related reasoning tasks, the reasoning capabilities of LLMs regarding other forms of reasoning are still an open problem. Here, we examine the issue of reasoning from the perspective of claim verification. We propose a framework designed to break down any claim paired with evidence into atomic reasoning types that are necessary for verification. We use this framework to create RECV, the first claim verification benchmark, incorporating real-world claims, to assess the deductive and abductive reasoning capabilities of LLMs. The benchmark comprises of three datasets, covering reasoning problems of in creasing complexity. We evaluate three state of-the-art proprietary LLMs under multiple prompt settings. Our results show that while LLMs can address deductive reasoning prob lems, they consistently fail in cases of abductive reasoning. Moreover, we observe that enhancing LLMs with rationale generation is not always beneficial. Nonetheless, we find that generated rationales are semantically similar to those provided by humans, especially in deduc tive reasoning cases.

RoleMRC: A Fine-Grained Composite Benchmark for Role-Playing and Instruction-Following

Junru Lu, Jiazheng Li, Guodong Shen, Lin Gui, Siyu An, Yulan He, di yin, Xing Sun

Role-playing is important for Large Language Models (LLMs) to follow diverse instructions while maintaining role identity and the role's pre-defined ability limits. Existing role-playing datasets mostly contribute to controlling role style and knowledge boundaries, but overlook role-playing in instruction-following scenarios. We introduce a fine-grained role-playing and instruction-following composite benchmark, named RoleMRC, including: (1) Multi-turn dialogues between ideal roles and humans, including free chats or discussions upon given passages; (2) Role-playing machine reading comprehension, involving response, refusal, and attempts according to passage answerability and role ability; (3) More complex scenarios with nested, multi-turn and prioritized instructions. The final RoleMRC features a 10.2k role profile meta-pool, 37.9k well-synthesized role-playing instructions, and 1.4k testing samples. We develop a pipeline to quantitatively evaluate the fine-grained role-playing and instruction-following capabilities of several mainstream LLMs, as well as models that are fine-tuned on our data. Moreover, cross-evaluation on external role-playing datasets confirms that models fine-tuned on RoleMRC enhances instruction-following without compromising general role-playing and reasoning capabilities. We also probe the neural-level activation maps of different capabilities over post-tuned LLMs. Access to our RoleMRC, RoleMRC-mix and Codes: https://github.com/LuJunru/RoleMRC.

Evaluation of Attribution Bias in Generator-Informed Retrieval-Augmented Large Language Models

Amin Abolghasemi, Leif Azzopardi, Seyyed Hadi Hashemi, Maarten de Rijke, Suzan Verberne

Attributing answers to source documents is an approach used to enhance the verifiability of a model's output in retrieval augmented generation (RAG). Prior work has mainly focused on improving and evaluating the attribution quality of large language models (LLMs) in RAG, but this may come at the expense of inducing biases in the attribution of answers. We define and examine two aspects in the evaluation of LLMs in RAG pipelines, namely attribution sensitivity and bias with respect to authorship information. We explicitly inform an LLM about the authors of source documents, instruct it to attribute its answers, and analyze (i) how sensitive the LLM's output is to the author of source documents, and (ii) whether the LLM exhibits a bias towards human-written or AI-generated source documents. We design an experimental setup in which we use counterfactual evaluation to study three LLMs in terms of their attribution sensitivity and bias in RAG pipelines. Our results show that adding authorship information to source documents can significantly change the attribution quality of LLMs by 3 to 18%. We show that LLMs can have an attribution bias towards explicit human authorship, which can serve as a competing hypothesis for findings of prior work that shows that LLM-generated content may be preferred over human-written contents. Our findings indicate that metadata of source documents can influence LLMs' trust, and how they attribute their answers. Furthermore, our research highlights attribution bias and sensitivity as a novel aspect of the vulnerability of LLMs.

Chumor: Towards Benchmarking Chinese Humor Understanding

Ruiqi He, Yushu He, Longju Bai, Jiarui Liu, Zhenjie Sun, Zenghao Tang, He Wang, Hanchen Xia, Rada Mihalcea, Naihao Deng

Existing humor datasets and evaluations predominantly focus on English, leaving limited resources for culturally nuanced humor in non-English languages like Chinese. To address this gap, we construct **Chumor**, the first Chinese humor explanation dataset that exceeds the size of existing humor datasets. **Chumor** is sourced from Ruo Zhi Ba (RZB,), a Chinese Reddit-like platform known for sharing intellectually challenging and culturally specific jokes. We test ten LLMs through both direct and chain-of-thought prompting, revealing that **Chumor** poses significant challenges to existing LLMs, with their accuracy slightly above random and far below human. In addition, our analysis highlights that human-annotated humor explanations are significantly better than explanations generated by both *GPT-4o* and *ERNIE*_{4\text{-}turbo}.

FanChuan: A Multilingual and Graph-Structured Benchmark For Parody Detection and Analysis

Yilun Zheng, Sha Li, Fangkun Wu, Yang Ziyi, Lin Hongchao, Zhichao Hu, Cai Xinjun, Ziming Wang, Jinxuan Chen, Sitao Luan, Jiahao Xu, Lihui Chen

Parody is an emerging phenomenon on social media, where individuals imitate a role or position opposite to their own, often for humor, provocation, or controversy. Detecting and analyzing parody can be challenging and is often reliant on context, yet it plays a crucial role in understanding cultural values, promoting subcultures, and enhancing self-expression. However, the study of parody is hindered by limited available data and deficient diversity in current datasets. To bridge this gap, we built seven parody datasets from both English and Chinese corpora, with 14,755 annotated users and 21,210 annotated comments in total. To provide sufficient context information, we also collect replies and construct user-interaction graphs to provide richer contextual information, which is lacking in existing datasets. With these datasets, we test traditional methods and Large Language Models (LLMs) on three key tasks: (1) parody detection, (2) comment sentiment analysis with parody, and (3) user sentiment analysis with parody. Our extensive experiments reveal that parody-related tasks still remain challenging for all models, and contextual information plays a critical role. Interestingly, we find that, in certain scenarios, traditional sentence embedding methods combined with simple classifiers can outperform advanced LLMs, i.e. DeepSeek-R1 and GPT-o3, highlighting parody as a significant challenge for LLMs.

MEDEC: A Benchmark for Medical Error Detection and Correction in Clinical Notes

Asma Ben Abacha, Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Meliha Yetisgen, Fei Xia, Thomas Lin

Several studies showed that Large Language Models (LLMs) can answer medical questions correctly, even outperforming the average human score in some medical exams. However, to our knowledge, no study has been conducted to assess the ability of language models to validate existing or generated medical text for correctness and consistency. In this paper, we introduce MEDEC, the first publicly available benchmark for medical error detection and correction in clinical notes, covering five types of errors (Diagnosis, Management, Treatment, Pharmacotherapy, and Causal Organism). MEDEC consists of 3,848 clinical texts, including 488 clinical notes from three US hospital systems that were not previously seen by any LLM. The dataset has been used in a shared task to evaluate seventeen participating systems. In this paper, we describe the data creation methods and we evaluate recent LLMs (e.g., o1-preview, GPT-4, Claude 3.5 Sonnet, Gemini 2.0 Flash, and DeepSeek-R1) for the tasks of detecting and correcting medical errors requiring both medical knowledge and reasoning capabilities. We also conducted a comparative study where two medical doctors performed the same task on the MEDEC test set. The results showed that MEDEC is a sufficiently challenging benchmark to assess the ability of models to validate existing or generated notes and to correct medical errors. We also found that although recent LLMs have a good performance in error detection and correction, they are still outperformed by medical doctors in these tasks. We discuss the potential factors behind this gap, the insights from our experiments, the limitations of current evaluation metrics, and share potential pointers for future research.

Financial Language Model Evaluation (FLaME)

Glenn Matlin, Mika Okamoto, Huzaifa Pardawala, Yang Yang, Sudheer Chava

Language Models (LMs) have demonstrated impressive capabilities with core Natural Language Processing (NLP) tasks. The effectiveness of LMs for highly specialized knowledge-intensive tasks in finance remains difficult to assess due to major gaps in the methodologies of existing evaluation frameworks, which have caused an erroneous belief in a far lower bound of LMs' performance on common Finance NLP (FinNLP) tasks. To demonstrate the potential of LMs for these FinNLP tasks, we present the first holistic benchmarking suite for Financial Language Model Evaluation (FLaME). We are the first research paper to comprehensively study LMs against 'reasoning-reinforced' LMs, with an empirical study of 23 foundation LMs over 20 core NLP tasks in finance. We open-source our framework software along with all data and results.

Verify with Caution: The Pitfalls of Relying on Imperfect Factuality Metrics

Ameya Godbole, Robin Jia

Improvements in large language models have led to increasing optimism that they can serve as reliable evaluators of natural language generation outputs. In this paper, we challenge this optimism by thoroughly re-evaluating five state-of-the-art factuality metrics on a collection of 11 datasets for summarization, retrieval-augmented generation, and question answering. We find that these evaluators are inconsistent with each other and often misestimate system-level performance, both of which can lead to a variety of pitfalls. We further show that these metrics exhibit biases against highly paraphrased outputs and outputs that draw upon faraway parts of the source documents. We urge users of these factuality metrics to proceed with caution and manually validate the reliability of these metrics in their domain of interest before proceeding.

TabXEval: Why this is a Bad Table? An eXhaustive Rubric for Table Evaluation

Vihang Pancholi, Jainit Sushil Bafna, Tejas Anvekar, Manish Shrivastava, Vivek Gupta

Evaluating reference tables qualitatively and quantitatively is a challenging problem, as traditional metrics often fail to capture the nuanced discrepancies in both structure and content. To mitigate this, we first introduce a methodical rubric that combines multi-level structural descriptors with fine-grained contextual quantification, laying a robust foundation for comprehensive table comparisons. Building on this, we propose **TabXEval**, an eXhaustive and eXplainable two-phase evaluation framework that initially aligns reference tables structurally via TabAlign and subsequently conducts a systematic semantic and syntactic comparison using TabCompare. This approach not only clarifies the evaluation process but also pinpoints subtle discrepancies that conventional methods overlook. To validate our framework, we present **TabXBench**, a diverse, multi-domain benchmark featuring realistic table perturbations and human-annotated assessments. Finally, we provide a systematic analysis of existing evaluation methods through sensitivity-specificity trade-offs and demonstrate the qualitative and quantitative effectiveness of **TabXEval** on text-to-table generation tasks, paving the way for future innovations in explainable table evaluation.

Evaluation of LLMs in Medical Text Summarization: The Role of Vocabulary Adaptation in High OOV Settings

Gunjan Balde, Soumyadeep Roy, Mainack Mondal, Niloy Ganguly

Large Language Models (LLMs) recently achieved great success in medical text summarization by simply using in-context learning. However, these recent efforts do not perform fine-grained evaluations under difficult settings where LLMs might fail. They typically report performance scores over the entire dataset. Through our benchmarking study, we show that LLMs show a significant performance drop for data points with high concentration of out-of-vocabulary (OOV) words or with high novelty. Vocabulary adaptation is an intuitive solution to this vocabulary mismatch issue where the LLM vocabulary gets updated with certain expert domain (here, medical) words or subwords. An interesting finding from our study is that Llama-3.1, even with a vocabulary size of around 128K tokens, still faces _over-fragmentation_ issue with medical words. To that end, we show vocabulary adaptation helps improve the LLM summarization performance even in difficult settings. Through extensive experimentation of multiple vocabulary adaptation strategies, two continual pretraining strategies, and three benchmark medical summarization datasets, we gain valuable insights into the role of vocabulary adaptation strategies for customizing LLMs to the medical domain. We also performed a human evaluation study with medical experts where they found that vocabulary adaptation results in more relevant and faithful summaries. Our codebase is made publicly available at https://github.com/gb-kgp/LLM-MedicalSummarization-Benchmark.

UniT: One Document, Many Revisions, Too Many Edit Intention Taxonomies

Fangping Lan, Abdullah Aljebreen, Eduard Dragut

Writing is inherently iterative, each revision enhancing information representation. One revision may contain many edits. Examination of the intentions behind edits provides valuable insights into an editor's expertise, the dynamics of collaborative writing, and the evolution of a document. Current research on edit intentions lacks a comprehensive edit intention taxonomy (EIT) that spans multiple application domains. As a result, researchers often create new EITs tailored to specific needs, a process that is both time-consuming and costly. To address this gap, we propose UniT, a Unified edit intention Taxonomy that integrates existing EITs encompassing a wide range of edit intentions. We examine the lineage relationship and the construction of 24 EITs. They together have 232 categories across various domains. During the literature survey and integration process, we identify challenges such as one-to-many category matches, incomplete definitions, and varying hierarchical structures. We propose solutions for resolving these issues. Finally, our evaluation shows that our UniT achieves higher inter-annotator agreement scores compared to existing EITs and is applicable to a large set of application domains.

DecompileBench: A Comprehensive Benchmark for Evaluating Decompilers in Real-World Scenarios

Zeyu Gao, Yuxin Cui, Hao Wang, Siliang Qin, Yuanda Wang, Zhang Bolun, Chao Zhang

Decompilers are fundamental tools for critical security tasks, from vulnerability discovery to malware analysis, yet their evaluation remains fragmented. Existing approaches primarily focus on syntactic correctness through synthetic micro-benchmarks or subjective human ratings, failing to address real-world requirements for semantic fidelity and analyst usability. We present **DecompileBench**, the first comprehensive framework that enables effective evaluation of decompilers in reverse engineering workflows through three key components: real-world function extraction (comprising 23,400 functions from 130 real-world programs), runtime-aware validation, and automated human-centric assessment using LLM-as-Judge to quantify the effectiveness of decompilers in reverse engineering workflows. Through a systematic comparison between six industrial-strength decompilers and six recent LLM-powered approaches, we demonstrate that LLM-based methods surpass commercial tools in code understandability despite 52.2% lower functionality correctness. These findings highlight the potential of LLM-based approaches to transform human-centric reverse engineering. We open source **DecompileBench** to provide a framework to advance research on decompilers and assist security experts in making informed tool selections based on their specific requirements.

The Behavior Gap: Evaluating Zero-shot LLM Agents in Complex Task-Oriented Dialogs

Avinash Baidya, Kamalika Das, Xiang Gao

Large Language Model (LLM)-based agents have significantly impacted Task-Oriented Dialog Systems (TODS) but continue to face notable performance challenges, especially in zero-shot scenarios. While prior work has noted this performance gap, the behavioral factors driving the performance gap remain under-explored. This study proposes a comprehensive evaluation framework to quantify the behavior gap between AI agents and human experts, focusing on discrepancies in dialog acts, tool usage, and knowledge utilization. Our findings reveal that this behavior gap is a critical factor negatively impacting the performance of LLM agents. Notably, as task complexity increases, the behavior gap widens (correlation: 0.963), leading to a degradation of agent performance on complex task-oriented dialogs. For the most complex task in our study, even the GPT-4o-based agent exhibits low alignment with human behavior, with low F1 scores for dialog acts (0.464), excessive and often misaligned tool usage with a F1 score of 0.139, and ineffective usage of external knowledge. Reducing such behavior gaps leads to significant performance improvement (22.4% on average). This study highlights the importance of comprehensive behavioral evaluations and improved alignment strategies to enhance the effectiveness of LLM-based TODS in handling complex tasks.

REPRO-Bench: Can AI Agents Assess the Reproducibility of Social Science Papers?

Chuxuan Hu, Liyun Zhang, Yeji Lim, Aum Wadhwani, Austin Peters, Daniel Kang

Assessing the reproducibility of social science papers is essential for promoting rigor in research processes, but manual assessment is costly. With advancements in AI agents, we seek to evaluate their capability to automate this process. However, existing benchmarks for reproducing research papers (1) focus solely on reproducing results using given code and data without assessing their consistency with the paper, (2) oversimplify real-world scenarios, and (3) lack necessary diversity in data formats and programming languages. To address these issues, we introduce REPRO-Bench, a collection of 112 task instances, each representing a social science paper with a publicly available reproduction report. The agents are tasked with assessing the reproducibility of the paper based on the original paper PDF and the corresponding reproduction package. REPRO-Bench features end-to-end evaluation tasks with complexity comparable to real-world assessments on the reproducibility of social science papers. We evaluate three representative AI agents on REPRO-Bench, with the best-performing agent achieving an accuracy of only 21.4%. Building on our empirical analysis, we develop REPRO-Agent, which improves the highest accuracy achieved by existing agents by 71%. We conclude that more advanced AI agents should be developed to automate real-world reproducibility assessment. REPRO-Bench is publicly available as an anonymous dataset: https://huggingface.co/datasets/anonymous-submission-acl2025/REPRO-Bench. The evaluation code is hosted anonymously: https://anonymous.4open.science/r/REPRO-Bench-Code-2E98/.

SceneGram: Conceptualizing and Describing Tangrams in Scene Context

Simeon Junker, Sina Zarrieß

Research on reference and naming suggests that humans can come up with very different ways of conceptualizing and referring to the same object, e.g. the same abstract tangram shape can be a "crab", "sink" or "space ship". Another common assumption in cognitive science is that scene context fundamentally shapes our visual perception of objects and conceptual expectations. This paper contributes SceneGram, a dataset of human references to tangram shapes placed in different scene contexts, allowing for systematic analyses of the effect of scene context on conceptualization. Based on this data, we analyze references to tangram shapes generated by multimodal LLMs, showing that these models do not account for the richness and variability of conceptualizations found in human references.

Compositional Syntactico-SemBanking for English as a Second or Foreign Language

Wenxi Li, Xihao Wang, Weiwei Sun

Despite the widespread use of English as a Second or Foreign Language (ESFL), developing syntactico-semantic representations for it is limited --- the irregularities in ESFL complicate systematic composition and subsequently the derivation of its semantics. This paper draws on constructivism and proposes a novel Synchronous Hyperedge Replacement Grammar (SHRG)-based constructivist approach to address the challenges. By using constructions as fundamental units, this approach not only accommodates both the idiosyncrasies and the compositional nature of ESFL, but also bridges the gap between literal cues and intended meaning. The feasibility of this constructivist approach is demonstrated using real ESFL data, resulting in a gold-standard, medium-sized syntactico-semantic bank that covers a wide range of ESFL phenomena. Experimental results further indicate that this newly developed resource significantly enhances syntactic parsing for ESFL, achieving an F_1 score improvement of over 20%.

Retrieval Models Aren't Tool-Savvy: Benchmarking Tool Retrieval for Large Language Models

Zhengliang Shi, Yuhan Wang, Lingyong Yan, Pengjie Ren, Shuaiqiang Wang, Dawei Yin, Zhaochun Ren

Tool learning aims to augment large language models (LLMs) with diverse tools, enabling them to act as agents for solving practical tasks. Due to the limited context length of tool-using LLMs, adopting information retrieval (IR) models to select useful tools from large toolsets is a critical initial step. However, the performance of IR models in tool retrieval tasks remains underexplored and unclear. Most tool-use benchmarks simplify this step by manually pre-annotating a small set of relevant tools for each task, which is far from the real-world scenarios. In this paper, we propose ToolRet, a heterogeneous tool retrieval benchmark comprising 7.6k diverse retrieval tasks, and a corpus of 43k tools, collected from existing datasets. We benchmark six types of models on ToolRet. Surprisingly, even the models with strong performance in conventional IR benchmarks, exhibit poor performance on ToolRet. This low retrieval quality degrades the task pass rate of tool-use LLMs. As a further step, we contribute a large-scale training dataset with over 200k instances, which substantially optimizes the tool retrieval ability of IR models.

TituLLMs: A Family of Bangla LLMs with Comprehensive Benchmarking

Shahriar Kabir Nahin, Rabindra Nath Nandi, Sagor Sarker, Quazi Sarwar Muhtaseem, Md Kowsher, Apu Chandraw Shill, Md Ibrahim, Mehadi Hasan Menon, Tareq Al Muntasir, Firoj Alam

In this paper, we present TituLLMs, the first large pretrained Bangla LLMs, available in 1b and 3b parameter sizes. Due to computational constraints during both training and inference, we focused on smaller models. To train TituLLMs, we collected a pretraining dataset of approximately 37 billion tokens. We extended the Llama-3.2 tokenizer to incorporate language- and culture-specific knowledge, which also enables faster training and inference. There was a lack of benchmarking datasets to benchmark LLMs for Bangla. To address this gap, we developed five benchmarking datasets. We benchmarked various LLMs, including TituLLMs, and demonstrated that TituLLMs outperforms its initial multilingual versions. However, this is not always the case, highlighting the complexities of language adaptation. Our work lays the groundwork for adapting existing multilingual open models to other low-resource languages. To facilitate broader adoption and further research, we have made the TituLLMs models and benchmarking datasets publicly available.

WikiMixQA: A Multimodal Benchmark for Question Answering over Tables and Charts

Negar Foroutan, Angelika Romanou, Matin Ansaripour, Julian Martin Eisenschlos, Karl Aberer, Rémi Lebret

Documents are fundamental to preserving and disseminating information, often incorporating complex layouts, tables, and charts that pose significant challenges for automatic document understanding (DU). While vision-language large models (VLLMs) have demonstrated improvements across various tasks, their effectiveness in processing long-context vision inputs remains unclear. This paper introduces WikiMixQA, a benchmark comprising 1,000 multiple-choice questions (MCQs) designed to evaluate cross-modal reasoning over tables and charts extracted from 4,000 Wikipedia pages spanning seven distinct topics. Unlike existing benchmarks, WikiMixQA emphasizes complex reasoning by requiring models to synthesize information from multiple modalities. We evaluate 12 state-of-the-art vision-language models, revealing that while proprietary models achieve 70% accuracy when provided with direct context, their performance deteriorates significantly when retrieval from long documents is required. Among these, GPT-4-o is the only model exceeding 50% accuracy in this setting, whereas open-source models perform considerably worse, with a maximum accuracy of 27%. These findings underscore the challenges of long-context, multi-modal reasoning and establish WikiMixQA as a crucial benchmark for advancing document understanding research.

The Million Authors Corpus: A Cross-Lingual and Cross-Domain Wikipedia Dataset for Authorship Verification

Abraham Israeli, Shuai Liu, Jonathan May, David Jurgens

Authorship verification (AV) is a crucial task for applications like identity verification, plagiarism detection, and AI-generated text identification. However, datasets for training and evaluating AV models are primarily in English and primarily in a single domain. This precludes analysis of AV techniques for generalizability and can cause seemingly valid AV solutions to, in fact, rely on topic-based features rather than actual authorship features. To address this limitation, we introduce the Million Authors Corpus (), a novel dataset encompassing contributions from dozens of languages on Wikipedia. It includes only long and contiguous textual chunks taken from Wikipedia edits and links those texts to their authors. includes 60.08M textual chunks, contributed by 1.29M Wikipedia authors. It enables broad-scale cross-lingual and cross-domain AV evaluation to ensure accurate analysis of model capabilities that are not overly optimistic. We provide baseline evaluations using state-of-the-art AV models as well as information retrieval models that are not AV-specific in order to demonstrate 's unique cross-lingual and cross-domain ablation capabilities.

Text2World: Benchmarking Large Language Models for Symbolic World Model Generation

Mengkang Hu, Tianxing Chen, Yude Zou, Yuheng Lei, Qiguang Chen, Ming Li, Yao Mu, Hongyuan Zhang, Wenqi Shao, Ping Luo

Recently, there has been growing interest in leveraging large language models (LLMs) to generate symbolic world models from textual descriptions. Although LLMs have been extensively explored in the context of world modeling, prior studies encountered several challenges, including evaluation randomness, dependence on indirect metrics, and a limited domain scope. To address these limitations, we introduce a novel benchmark, Text2World, based on planning domain definition language (PDDL), featuring hundreds of diverse domains and employing multi-criteria, execution-based metrics for a more robust evaluation. We benchmark current LLMs using Text2World and find that reasoning models trained with large-scale reinforcement learning outperform others. However, even the best-performing model still demonstrates limited capabilities in world modeling. Building on these insights, we examine several promising strategies to enhance the world modeling capabilities of LLMs, including test-time scaling, agent training, and more. We hope that Text2World can serve as a crucial resource, laying the groundwork for future research in leveraging LLMs as world models.

T^5Score: A Methodology for Automatically Assessing the Quality of LLM Generated Multi-Document Topic Sets

Itamar Trainin, Omri Abend

Using LLMs for Multi-Document Topic Extraction has recently gained popularity due to their apparent high-quality outputs, expressiveness, and ease of use. However, most existing evaluation practices are not designed for LLM-generated topics and result in low inter-annotator agreement scores, hindering the reliable use of LLMs for the task. To address this, we introduce T^5Score, an evaluation methodology that decomposes the quality of a topic set into quantifiable aspects, measurable through easy-to-perform annotation tasks. This framing enables a convenient, manual or automatic, evaluation procedure resulting in a strong inter-annotator agreement score. To substantiate our methodology and claims, we perform extensive experimentation on multiple datasets and report the results.

LLM-as-an-Interviewer: Beyond Static Testing Through Dynamic LLM Evaluation

Eunsu Kim, Juyoung Suk, Seungone Kim, Niklas Muennighoff, Dongkwan Kim, Alice Oh

We introduce LLM-as-an-Interviewer, a novel paradigm for evaluating large language models (LLMs). This approach leverages multi-turn interactions where the LLM interviewer actively provides feedback on responses and poses follow-up questions to the evaluated LLM. At the start of the interview, the LLM interviewer dynamically modifies datasets to generate initial questions, mitigating data contamination. We apply the LLM-as-an-Interviewer framework to evaluate six models on the reasoning, factuality and instruction-following tasks. Our results show that the framework effectively provides insights into LLM performance, including the quality of initial responses, adaptability to feedback, and ability to address follow-up queries like clarification or additional knowledge requests. The framework also addresses key limitations of conventional methods like LLM-as-a-Judge, including verbosity bias and inconsistency across runs. Finally, we propose the Interview Report, which aggregates insights from the interview process, providing examples and a comprehensive analysis of the LLM's strengths and weaknesses. This report offers a detailed snapshot of the model's real-world applicability.

NBDESCRIB: A Dataset for Text Description Generation from Tables and Code in Jupyter Notebooks with Guidelines

Xuye Liu, Tengfei Ma, Yimu Wang, Fengjie Wang, Jian Zhao

Generating cell-level descriptions for Jupyter Notebooks, which is a major resource consisting of codes, tables, and descriptions, has been attracting increasing research attention. However, existing methods for Jupyter Notebooks mostly focus on generating descriptions from code snippets or table outputs independently. On the other side, descriptions should be personalized as users have different purposes in different scenarios while previous work ignored this situation during description generation. In this work, we formulate a new task, personalized description generation with code, tables,and user-written guidelines in Jupyter Notebooks. To evaluate this new task, we collect and propose a benchmark, namely NBDESCRIB, containing code, tables, and user-written guidelines as inputs and personalized descriptions as targets. Extensive experiments show that while existing models of text generation are able to generate fluent and readable descriptions, they still struggle to produce factually correct descriptions without user-written guidelines. CodeT5 achieved the highest scores in Orientation (1.27) and Correctness (-0.43) among foundation models in human evaluation, while the ground truth scored higher in Orientation (1.45) and Correctness (1.19). Common error patterns involve misalignment with guidelines, incorrect variable values, omission of im-031 portant code information, and reasoning errors.032 Moreover, ablation studies show that adding guidelines significantly enhances performance, both qualitatively and quantitatively.

Standard Quality Criteria Derived from Current NLP Evaluations for Guiding Evaluation Design and Grounding Comparability and AI Compliance Assessments

Anya Belz, Simon Mille, Craig Thomson

Research shows that two evaluation experiments reporting results for the same quality criterion name (e.g. Fluency) do not necessarily evaluate the same aspect of quality. Not knowing when two evaluations are comparable in this sense means we currently lack the ability to draw conclusions based on multiple independently conducted evaluations. It is hard to see how this issue can be fully addressed other than by the creation of a standard set of quality criterion names and definitions that the evaluations in use in NLP can be grounded in. Taking a descriptivist approach, the QCET Quality Criteria for Evaluation Taxonomy derives a standard set of quality criterion names and definitions from two surveys of evaluations in NLP, and structures them into a reference taxonomy. We present QCET and its uses in (i) establishing comparability of existing evalua- tions, (ii) guiding the design of new evaluations, and (iii) assessing regulation compliance.

skLEP: A Slovak General Language Understanding Benchmark

Marek Suppa, Andrej Ridzik, Daniel Hládek, Tomáš Javůrek, Viktória Ondrejová, Kristína Sásiková, Martin Tamajka, Marian Simko

This paper introduces skLEP, a comprehensive benchmark for evaluating Slovak natural language understanding (NLU) models. The benchmark comprises nine diverse tasks spanning token-level, sentence-pair, and document-level challenges. To create skLEP, we curated new datasets and translated established English resources. The paper presents extensive baseline evaluations of various Slovak-specific, multilingual, and English models. Additionally, we release an open-source toolkit and a standardized leaderboard to foster reproducibility and drive future research in Slovak NLU.

MMRefine: Unveiling the Obstacles to Robust Refinement in Multimodal Large Language Models

Gio Paik, Geewook Kim, Jinbae Im

This paper introduces MMRefine, a Multi-Modal Refinement Benchmark designed to evaluate the error refinement capabilities of Multimodal Large Language Models (MLLMs). As the emphasis shifts toward enhancing reasoning during inference, MMRefine provides a framework that assesses refinement beyond final answer accuracy by categorizing errors into diverse levels. The benchmark evaluates MLLMs' abilities to detect and correct errors across six distinct scenarios. Experiments with various open and closed MLLMs reveal refinement bottlenecks and factors impeding performance, highlighting areas for improvement in effective reasoning enhancement.

DynaQuest: A Dynamic Question Answering Dataset Reflecting Real-World Knowledge Updates

Qian Lin, Junyi Li, Hwee Tou Ng

The rapidly changing nature of real-world information presents challenges for large language models (LLMs), which are typically trained on static datasets. This limitation makes it difficult for LLMs to accurately perform tasks that require up-to-date knowledge, such as time-sensitive question answering (QA). In this paper, we introduce **DynaQuest**, a **Dyna**mic **Quest**ion answering dataset reflecting knowledge updates in the real world. DynaQuest is based on Wikipedia Infoboxes, which are frequently updated to reflect real-world changes. Our dataset is created by automatically identifying and comparing changes between different versions of Wikipedia pages and generating question-answer pairs based on these updates. To address the challenges posed by our dynamic dataset, we propose **CARL**, a **C**ontext-**A**ware **R**einforcement **L**earning framework to improve the performance of LLMs on time-sensitive question answering. We conduct experiments on our collected dataset across recent time periods and demonstrate the effectiveness of our approach. Furthermore, we maintain a dynamic knowledge updating process, providing a periodically evolving benchmark to continually evaluate LLMs' ability to answer time-sensitive questions.

ELABORATION: A Comprehensive Benchmark on Human-LLM Competitive Programming

Xinwei Yang, Zhaofeng Liu, Chen Huang, Jiashuai Zhang, Tong Zhang, Yifan Zhang, Wenqiang Lei

While recent research increasingly emphasizes the value of human-LLM collaboration in competitive programming and proposes numerous empirical methods, a comprehensive understanding remains elusive due to the fragmented nature of existing studies and their use of diverse, application-specific human feedback. Thus, our work serves a three-fold purpose: First, we present the first taxonomy of human feedback consolidating the entire programming process, which promotes fine-grained evaluation. Second, we introduce ELABORATIONSET, a novel programming dataset specifically designed for human-LLM collaboration, meticulously annotated to enable large-scale simulated human feedback and facilitate cost-effective real human interaction studies. Third, we introduce ELABORATION, a novel benchmark to facilitate a thorough assessment of human-LLM competitive programming. With ELABORATION, we pinpoint strengthes and weaknesses of existing methods, thereby setting the foundation for furture improvement. Our dataset and code will be openly released.

StrucText-Eval: Evaluating Large Language Model's Reasoning Ability in Structure-Rich Text

Zhouhong Gu, Haoning Ye, Xingzhou Chen, Zeyang Zhou, Hongwei Feng, Yanghua Xiao

The effective utilization of structured data, integral to corporate data strategies, has been challenged by the rise of large language models (LLMs) capable of processing unstructured information. This shift prompts the question: can LLMs interpret structured data directly in its unstructured form? We propose an automatic evaluation data generation method for assessing LLMs' reasoning capabilities on structure-rich text to explore this. Our approach supports 8 structured languages and 29 tasks, generating data with adjustable complexity through controllable nesting and structural width. We introduce StrucText-Eval, a benchmark containing 5,800 pre-generated and annotated samples designed to evaluate how well LLMs understand and reason through structured text. StrucText-Eval is divided into two suites: a regular Test suite (3,712 samples) and a Test-Hard suite (2,088 samples), the latter emphasizing the gap between human and model performance on more complex tasks. Experimental results show that while open-source LLMs achieve a maximum accuracy of 74.9% on the standard dataset, their performance drops significantly to 45.8% on the harder dataset. In contrast, human participants reach an accuracy of 92.6% on StrucText-Eval-Hard, highlighting LLMs' current limitations in handling intricate structural information.

Evaluating Lexical Proficiency in Neural Language Models

Cristiano Ciaccio, Alessio Miaschi, Felice Dell'Orletta

We present a novel evaluation framework designed to assess the lexical proficiency and linguistic creativity of Transformer-based Language Models (LMs). We validate the framework by analyzing the performance of a set of LMs of different sizes, in both mono- and multilingual configuration, across tasks involving the generation, definition, and contextual usage of lexicalized words, neologisms, and nonce words. To support these evaluations, we developed a novel dataset of lexical entries for the Italian language, including curated definitions and usage examples sourced from various online platforms. The results highlight the robustness and effectiveness of our framework in evaluating multiple dimensions of LMs' linguistic understanding and offer an insight, through the assessment of their linguistic creativity, on the lexical generalization abilities of LMs.

Real-time Fake News from Adversarial Feedback

Sanxing Chen, Yukun Huang, Bhuwan Dhingra

We show that existing evaluations for assessing the factuality of news from conventional sources, such as claims on fact-checking websites, result in high accuracies over time for LLM-based detectors—even after their knowledge cutoffs. This suggests that recent popular false information from such sources can be easily identified due to its likely presence in pre-training/retrieval corpora or the emergence of salient, yet shallow, patterns in these datasets. Instead, we argue that a proper factuality evaluation dataset should test a model's ability to reason about current events by retrieving and reading related evidence. To this end, we develop a novel pipeline that leverages natural language feedback from a RAG-based detector to iteratively modify real-time news into deceptive variants that challenge LLMs. Our iterative rewrite decreases the binary classification ROC-AUC by an absolute 17.5 percent for a strong RAG-based GPT-4o detector. Our experiments reveal the important role of RAG in both evaluating and generating challenging news examples, as retrieval-free LLM detectors are vulnerable to unseen events and adversarial attacks, while feedback from RAG-based evaluation helps discover more deceitful patterns.

BQA: Body Language Question Answering Dataset for Video Large Language Models

Shintaro Ozaki, Kazuki Hayashi, Miyu Oba, Yusuke Sakai, Hidetaka Kamigaito, Taro Watanabe

A large part of human communication relies on nonverbal cues such as facial expressions, eye contact, and body language. Unlike language or sign language, such nonverbal communication lacks formal rules, requiring complex reasoning based on commonsense understanding. Enabling current Video Large Language Models (VideoLLMs) to accurately interpret body language is a crucial challenge, as human unconscious actions can easily cause the model to misinterpret their intent. To address this, we propose a dataset, BQA, a body language question answering dataset, to validate whether the model can correctly interpret emotions from short clips of body language comprising 26 emotion labels of videos of body language. We evaluated various VideoLLMs on the BQA with and without Multimodal Chain of Thought (CoT) and revealed that understanding body language is challenging, and our analyses of the wrong answers by VideoLLMs show that certain VideoLLMs made largely biased answers depending on the age group and ethnicity of the individuals. We also found consistent error patterns in VideoLLMs.

UTBoost: Rigorous Evaluation of Coding Agents on SWE-Bench

Boxi Yu, Yuxuan Zhu, Pinjia He, Daniel Kang

The advent of Large Language Models (LLMs) has spurred the development of coding agents for real-world code generation. As a widely used benchmark for evaluating the code generation capabilities of these agents, SWE-Bench uses real-world problems based on GitHub issues and their corresponding pull requests. However, the manually written test cases included in these pull requests are often insufficient, allowing generated patches to pass the tests without resolving the underlying issue. To address this challenge, we introduce UTGenerator, an LLM-driven test case generator that automatically analyzes codebases and dependencies to generate test cases for real-world Python projects. Building on UTGenerator, we propose UTBoost, a comprehensive framework for test case augmentation. In our evaluation, we identified 36 task instances with insufficient test cases and uncovered 345 erroneous patches incorrectly labeled as passed in the original SWE Bench. These corrections, impacting 40.9\% of SWE-Bench Lite and 24.4\% of SWE-Bench Verified leaderboard entries, yield 18 and 11 ranking changes, respectively.

RAGEval: Scenario Specific RAG Evaluation Dataset Generation Framework

Kunlun Zhu, Yifan Luo, Dingling Xu, Yukun Yan, Zhenghao Liu, Shi Yu, Ruobing Wang, Shuo Wang, Yishan Li, Nan Zhang, Xu Han, Zhiyuan Liu, Maosong Sun

Retrieval-Augmented Generation (RAG) is a powerful approach that enables large language models (LLMs) to incorporate external knowledge. However, evaluating the effectiveness of RAG systems in specialized scenarios remains challenging due to the high costs of data construction and the lack of suitable evaluation metrics. This paper introduces RAGEval, a framework designed to assess RAG systems across diverse scenarios by generating high-quality documents, questions, answers, and references through a schema-based pipeline. With a focus on factual accuracy, we propose three novel metrics—Completeness, Hallucination, and Irrelevance—to evaluate LLM-generated responses rigorously. Experimental results show that RAGEval outperforms zero-shot and one-shot methods in terms of clarity, safety, conformity, and richness of generated samples. Furthermore, the use of LLMs for scoring the proposed metrics demonstrates a high level of consistency with human evaluations. RAGEval establishes a new paradigm for evaluating RAG systems in real-world applications. The code and dataset are released at https://github.com/OpenBMB/RAGEval.

Sinhala Encoder-only Language Models and Evaluation

Tharindu Ranasinghe, Hansi Hettiarachchi, Nadeesha Chathurangi Naradde Vidana Pathirana, Damith Premasiri, Lasitha Uyangodage, Isuri Nanomi Arachchige, Alistair Plum, Paul Rayson, Ruslan Mitkov

Recently, language models (LMs) have produced excellent results in many natural language processing (NLP) tasks. However, their effectiveness is highly dependent on available pre-training resources, which is particularly challenging for low-resource languages such as Sinhala. Furthermore, the scarcity of benchmarks to evaluate LMs is also a major concern for low-resource languages. In this paper, we address these two challenges for Sinhala by (i) collecting the largest monolingual corpus for Sinhala, (ii) training multiple LMs on this corpus and (iii) compiling the first Sinhala NLP benchmark (Sinhala-GLUE) and evaluating LMs on it. We show the Sinhala LMs trained in this paper outperform the popular multilingual LMs, such as XLM-R and existing Sinhala LMs in downstream NLP tasks. All the trained LMs are publicly available. We also make Sinhala-GLUE publicly available as a public leaderboard, and we hope that it will enable further advancements in developing and evaluating LMs for Sinhala.

Pragmatics in the Era of Large Language Models: A Survey on Datasets, Evaluation, Opportunities and Challenges

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, Barbara Plank

Understanding pragmatics—the use of language in context—is crucial for developing NLP systems capable of interpreting nuanced language use. Despite recent advances in language technologies, including large language models, evaluating their ability to handle pragmatic phenomena such as implicatures and references remains challenging. To advance pragmatic abilities in models, it is essential to understand current evaluation trends and identify existing limitations. In this survey, we provide a comprehensive review of resources designed for evaluating pragmatic capabilities in NLP, categorizing datasets by the pragmatic phenomena they address. We analyze task designs, data collection methods, evaluation approaches, and their relevance to real-world applications. By examining these resources in the context of modern language models, we highlight emerging trends, challenges, and gaps in existing benchmarks. Our survey aims to clarify the landscape of pragmatic evaluation and guide the development of more comprehensive and targeted benchmarks, ultimately contributing to more nuanced and context-aware NLP models.

VMLU Benchmarks: A comprehensive benchmark toolkit for Vietnamese LLMs

Cuc Thi Bui, Nguyen Truong Son, Truong van trang, Lam Viet Phung, Pham Nhut Huy, Hoang Anh Le, Quoc Huu Van, Phong Nguyen-Thuan Do, Van Le Tran Truc, Duc Thanh Chau, Le-Minh Nguyen

The evolution of Large Language Models (LLMs) has underscored the necessity for benchmarks designed for various languages and cultural contexts. To address this need for Vietnamese, we present the first Vietnamese Multitask Language Understanding (VMLU) Benchmarks. The VMLU benchmarks consist of four datasets that assess different capabilities of LLMs, including general knowledge, reading comprehension, reasoning, and conversational skills. This paper also provides an insightful overview of the current state of some dominant LLMs, such as Llama-3, Qwen2.5, and GPT-4, highlighting their performances and limitations when measured against these benchmarks. Furthermore, we provide insights into how prompt design can influence VMLU's evaluation outcomes, as well as suggest that open-source LLMs can serve as effective, cost-efficient evaluators within the Vietnamese context. By offering a comprehensive and accessible benchmarking framework, the VMLU Benchmarks aim to foster the development and fine-tuning of Vietnamese LLMs, thereby establishing a foundation for their practical applications in language-specific domains.

Beyond Surface Simplicity: Revealing Hidden Reasoning Attributes for Precise Commonsense Diagnosis

Huijun Lian, Zekai Sun, Keqi Chen, Yingming Gao, Ya Li

Commonsense question answering (QA) are widely used to evaluate the commonsense abilities of large language models. However, answering commonsense questions correctly requires not only knowledge but also reasoning—even for seemingly simple questions. We demonstrate that such hidden reasoning attributes in commonsense questions can lead evaluation accuracy differences of up to 24.8% across different difficulty levels in the same benchmark. Current benchmarks overlook these hidden reasoning attributes, making it difficult to assess a model's specific levels of commonsense knowledge and reasoning ability. To address this issue, we introduce ReComSBench, a novel framework that reveals hidden reasoning attributes behind commonsense questions by leveraging the knowledge generated during the reasoning process. Additionally, ReComSBench proposes three new metrics for decoupled evaluation: Knowledge Balanced Accuracy, Marginal Sampling Gain, and Knowledge Coverage Ratio. Experiments show that ReComSBench provides insights into model performance that traditional benchmarks cannot offer. The difficulty stratification based on revealed hidden reasoning attributes performs as effectively as the model-probability-based approach but is more generalizable and better suited for improving a model's commonsense reasoning abilities. By uncovering and analyzing the hidden reasoning attributes in commonsense data, ReComSBench offers a new approach to enhancing existing commonsense benchmarks.

CRiskEval: A Chinese Multi-Level Risk Evaluation Benchmark Dataset for Large Language Models

Ling Shi, Deyi Xiong

Large language models (LLMs) are possessed of numerous beneficial capabilities, yet their potential inclination harbors unpredictable risks that may materialize in the future. We hence propose CRiskEval, a Chinese dataset meticulously designed for gauging the risk proclivities inherent in LLMs such as resource acquisition and malicious coordination, as part of efforts for proactive preparedness. To curate CRiskEval, we define a new risk taxonomy with 7 types of frontier risks and 4 safety levels, including extremely hazardous,moderately hazardous, neutral and safe. We follow the philosophy of tendency evaluation to empirically measure the stated "desire" of LLMs via fine-grained multiple-choice question answering. The dataset consists of 14,888 questions that simulate scenarios related to predefined 7 types of frontier risks. Each question is accompanied with 4 answer choices that state opinions or behavioral tendencies corresponding to the question. All answer choices are manually annotated with one of the defined risk levels so that we can easily build a fine-grained frontier risk profile for each assessed LLM. Extensive evaluation with CRiskEval on a spectrum of prevalent Chinese LLMs has unveiled a striking revelation: most models exhibit risk tendencies of more than 40% (weighted tendency to the four risk levels). Furthermore, a subtle increase in the model's inclination toward urgent self-sustainability, power seeking and other dangerous goals becomes evident as the size of models increases. To promote further research on the frontier risk evaluation of LLMs, we publicly release our dataset at https://github.com/tjunlp-lab/CRiskEval.

Demystifying Small Language Models for Edge Deployment

Zhenyan Lu, Xiang Li, Dongqi Cai, Rongjie Yi, Fangming Liu, Wei Liu, Jian Luan, Xiwen Zhang, Nicholas D. Lane, Mengwei Xu

Small language models (SLMs) have emerged as a promising solution for deploying resource-constrained devices, such as smartphones and Web of Things. This work presents the first comprehensive study of over 60 SLMs such as Microsoft Phi and Google Gemma that are publicly accessible. Our findings show that state-of-the-art SLMs outperform 7B models in general tasks, proving their practical viability. However, SLMs' in-context learning capabilities remain limited, and their efficiency has significant optimization potential. We identify key SLM optimization opportunities, including dynamic task-specific routing, model-hardware co-design, and vocabulary/KV cache compression. Overall, we expect the work to reveal an all-sided landscape of SLMs, benefiting the research community across algorithm, model, system, and hardware levels.

SwiLTra-Bench: The Swiss Legal Translation Benchmark

Joel Niklaus, Jakob Merane, Luka Nenadic, Sina Ahmadi, Yingqiang Gao, Cyrill A. H. Chevalley, Claude Humbel, Christophe Gösken, Lorenzo Tanzi, Thomas Lüthi, Stefan Palombo, Spencer Poff, Boling Yang, Nan Wu, Matthew Guillod, Robin Mamié, Daniel Brunner, Julio Pereyra, Niko Grupen

In Switzerland legal translation is uniquely important due to the country's four official languages and requirements for multilingual legal documentation. However, this process traditionally relies on professionals who must be both legal experts and skilled translators---creating bottlenecks and impacting effective access to justice. To address this challenge, we introduce SwiLTra-Bench, a comprehensive multilingual benchmark of over 180K aligned Swiss legal translation pairs comprising laws, headnotes, and press releases across all Swiss languages along with English, designed to evaluate LLM-based translation systems. Our systematic evaluation reveals that frontier models achieve superior translation performance across all document types, while specialized translation systems excel specifically in laws but under-perform in headnotes. Through rigorous testing and human expert validation, we demonstrate that while fine-tuning open SLMs significantly improves their translation quality, they still lag behind the best zero-shot prompted frontier models such as Claude-3.5-Sonnet. Additionally, we present SwiLTra-Judge, a specialized LLM evaluation system that aligns best with human expert assessments.

ProxAnn: Use-Oriented Evaluations of Topic Models and Document Clustering

Alexander Miserlis Hoyle, Lorena Calvo-Bartolomé, Jordan Lee Boyd-Graber, Philip Resnik

Topic models and document-clustering evaluations either use automated metrics that align poorly with human preferences, or require expert labels that are intractable to scale. We design a scalable human evaluation protocol and a corresponding automated approximation that reflect practitioners' real-world usage of models. Annotators---or an LLM-based proxy---review text items assigned to a topic or cluster, infer a category for the group, then apply that category to other documents. Using this protocol, we collect extensive crowdworker annotations of outputs from a diverse set of topic models on two datasets. We then use these annotations to validate automated proxies, finding that the best LLM proxy is statistically indistinguishable from a human annotator and can therefore serve as a reasonable substitute in automated evaluations.

CrisisTS: Coupling Social Media Textual Data and Meteorological Time Series for Urgency Classification

Romain Meunier, Farah Benamara, Véronique Moriceau, Savitha Ramasamy, Zhongzheng Qiao

This paper proposes CrisisTS, the first multimodal and multilingual dataset for urgency classification composed of benchmark crisis datasets from French and English social media about various expected (e.g., flood, storm) and sudden (e.g., earthquakes, explosions) crises that have been mapped with open source geocoded meteorological time series data. This mapping is based on a simple and effective strategy that allows for temporal and location alignment even in the absence of location mention in the text. A set of multimodal experiments have been conducted relying on transformers and LLMs to improve overall performances while ensuring model generalizability. Our results show that modality fusion outperforms text-only models.

EvolveBench: A Comprehensive Benchmark for Assessing Temporal Awareness in LLMs on Evolving Knowledge

Zhiyuan Zhu, Yusheng Liao, Zhe Chen, Yuhao Wang, Yunfeng Guan, Yanfeng Wang, Yu Wang

Large language models (LLMs) are trained on extensive historical corpora, but their ability to understand time and maintain temporal awareness of time-evolving factual knowledge remains limited. Previous studies often neglect the critical aspect of utilizing knowledge from various sources. To address this gap, we introduce EvolveBench, a comprehensive benchmark that evaluates temporal competence along five key dimensions: Cognition, which examines the ability to recall and contextualize historical facts. Awareness, which tests LLMs' awareness of temporal misalignment between external inputs and the temporal context of a query. Trustworthiness, which assesses whether models can identify and appropriately refuse queries based on invalid timestamps. Understanding, which focuses on interpreting both explicit dates and implicit historical markers. Finally, reasoning evaluates the capacity to analyze temporal relationships and draw accurate inferences. Evaluating 15 widely used LLMs on EvolveBench shows that GPT-4o achieves the highest average EM score of 79.36, while the open-source Llama3.1-70B demonstrates notable strength in handling temporally misaligned contexts with an average score of 72.47. Despite these advances, all models still struggle with handling temporal misaligned context. Our code and dataset are available at https://github.com/zzysjtuiwct/EvolveBench.

What Matters in Evaluating Book-Length Stories? A Systematic Study of Long Story Evaluation

Dingyi Yang, Qin Jin

In this work, we conduct systematic research in a challenging area: the automatic evaluation of book-length stories (>100K tokens). Our study focuses on two key questions: (1) understanding which evaluation aspects matter most to readers, and (2) exploring effective methods for evaluating lengthy stories. We introduce the first large-scale benchmark, **LongStoryEval**, comprising 600 newly published books with an average length of 121K tokens (maximum 397K). Each book includes its average rating and multiple reader reviews, presented as critiques organized by evaluation aspects. By analyzing all user-mentioned aspects, we propose an *evaluation criteria structure* and conduct experiments to identify the most significant aspects among the 8 top-level criteria. For evaluation methods, we compare the effectiveness of three types: *aggregation-based*, *incremental-updated*, and *summary-based* evaluations. Our findings reveal that aggregation- and summary-based evaluations perform better, with the former excelling in detail assessment and the latter offering greater efficiency. Building on these insights, we further propose **NovelCritique**, an 8B model that leverages the efficient summary-based framework to review and score stories across specified aspects. NovelCritique outperforms commercial models like GPT-4o in aligning with human evaluations. All our datasets and codes will be released to foster further research.

LIFBench: Evaluating the Instruction Following Performance and Stability of Large Language Models in Long-Context Scenarios

Xiaodong Wu, Minhao Wang, Yichen Liu, Xiaoming Shi, He Yan, Lu Xiangju, Junmin Zhu, Wei Zhang

As Large Language Models (LLMs) evolve in natural language processing (NLP), their ability to stably follow instructions in long-context inputs has become critical for real-world applications. However, existing benchmarks seldom focus on instruction-following in long-context scenarios or stability on different inputs. To bridge this gap, we introduce LIFBench, a scalable dataset designed to evaluate LLMs' instruction-following capabilities and stability across long contexts. LIFBench comprises three long-context scenarios and eleven diverse tasks, featuring 2,766 instructions generated through an automated expansion method across three dimensions: length, expression, and variables. For evaluation, we propose LIFEval, a rubric-based assessment method that enables precise, automated scoring of complex LLM responses without reliance on LLM-assisted assessments or human judgment. This method allows for a comprehensive analysis of model performance and stability from multiple perspectives. We conduct detailed experiments on 20 prominent LLMs across six length intervals. Our work contributes LIFBench and LIFEval as robust tools for assessing LLM performance in complex and long-context settings, offering valuable insights to guide future advancements in LLM development.

PhysReason: A Comprehensive Benchmark towards Physics-Based Reasoning

Xinyu Zhang, Yuxuan Dong, Yanrui Wu, Jiaxing Huang, Chengyou Jia, Basura Fernando, Mike Zheng Shou, Lingling Zhang, Jun Liu

Large language models demonstrate remarkable capabilities across various domains, especially mathematics and logic reasoning. However, current evaluations overlook physics-based reasoning - a complex task requiring physics theorems and constraints. We present PhysReason, a 1,200-problem benchmark comprising knowledge-based (25%) and reasoning-based (75%) problems, where the latter are divided into three difficulty levels (easy, medium, hard). Notably, problems require an average of 8.1 solution steps, with hard requiring 15.6, reflecting the complexity of physics-based reasoning. We propose the Physics Solution Auto Scoring Framework, incorporating efficient answer-level and comprehensive step-level evaluations. Top-performing models like Deepseek-R1, Gemini-2.0-Flash-Thinking, and o3-mini-high achieve less than 60% on answer-level evaluation, with performance dropping from knowledge questions (75.11%) to hard problems (31.95%). Through step-level evaluation, we identified four key bottlenecks: Physics Theorem Application, Physics Process Understanding, Calculation, and Physics Condition Analysis. These findings position PhysReason as a novel and comprehensive benchmark for evaluating physics-based reasoning capabilities in large language models.

FEA-Bench: A Benchmark for Evaluating Repository-Level Code Generation for Feature Implementation

Wei Li, Xin Zhang, Zhongxin Guo, Shaoguang Mao, Wen Luo, Guangyue Peng, Yangyu Huang, Houfeng Wang, Scarlett Li

Implementing new features in repository-level codebases is a crucial application of code generation models. However, current benchmarks lack a dedicated evaluation framework for this capability. To fill this gap, we introduce FEA-Bench, a benchmark designed to assess the ability of large language models (LLMs) to perform incremental development within code repositories. We collect pull requests from 83 GitHub repositories and use rule-based and intent-based filtering to construct task instances focused on new feature development. Each task instance containing code changes is paired with relevant unit test files to ensure that the solution can be verified. The feature implementation requires LLMs to simultaneously possess code completion capabilities for new components and code editing abilities for other relevant parts in the code repository, providing a more comprehensive evaluation method of LLMs' automated software engineering capabilities. Experimental results show that LLMs perform significantly worse in the FEA-Bench, highlighting considerable challenges in such repository-level incremental code development.

Can You Really Trust Code Copilot? Evaluating Large Language Models from a Code Security Perspective

Yutao Mou, Xiao Deng, Yuxiao Luo, Shikun Zhang, Wei Ye

Code security and usability are both essential for various coding assistant applications driven by large language models (LLMs). Current code security benchmarks focus solely on single evaluation task and paradigm, such as code completion and generation, lacking comprehensive assessment across dimensions like secure code generation, vulnerability repair and discrimination. In this paper, we first propose CoV-Eval, a multi-task benchmark covering various tasks such as code completion, vulnerability repair, vulnerability detection and classification, for comprehensive evaluation of LLM code security. Besides, we developed VC-Judge, an improved judgment model that aligns closely with human experts and can review LLM-generated programs for vulnerabilities in a more efficient and reliable way. We conduct a comprehensive evaluation of 20 proprietary and open-source LLMs. Overall, while most LLMs identify vulnerable codes well, they still tend to generate insecure codes and struggle with recognizing specific vulnerability types and performing repairs. Extensive experiments and qualitative analyses reveal key challenges and optimization directions, offering insights for future research in LLM code security.

An Expanded Massive Multilingual Dataset for High-Performance Language Technologies

Laurie Burchell, Ona De Gibert Bonet, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Pinzhen Chen, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Jindřich Helcl, Erik Henriksson, Mateusz Klimaszewski, Ville Komulainen, Andrey Kutuzov, Joona Kytöniemi, Veronika Laippala, Petter Mæhlum, Bhavitvya Malik, Farrokh Mehryary, Vladislav Mikhailov, Nikita Moghe, Amanda Myntti, Dayyán O'Brien, Stephan Oepen, Proyag Pal, Jousia Piha, Sampo Pyysalo, Gema Ramírez-Sánchez, David Samuel, Pavel Stepachev, Jörg Tiedemann, Dušan Variš, Tereza Vojtěchová, Jaume Zaragoza-Bernabeu

Training state-of-the-art large language models requires vast amounts of clean and diverse textual data. However, building suitable multilingual datasets remains a challenge. In this work, we present HPLT v2, a collection of high-quality multilingual monolingual and parallel corpora, extending prior work of the HPLT project. The monolingual portion of the data contains 8T tokens covering 193 languages, while the parallel data contains 380M sentence pairs covering 51 languages. We document the entire data pipeline and release the code to reproduce it. We provide extensive analysis of the quality and characteristics of our data. Finally, we evaluate the performance of language models and machine translation systems trained on HPLT v2, demonstrating its value.

Agent-RewardBench: Towards a Unified Benchmark for Reward Modeling across Perception, Planning, and Safety in Real-World Multimodal Agents

Tianyi Men, Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao

As Multimodal Large Language Models (MLLMs) advance, multimodal agents show promise in real-world tasks like web navigation and embodied intelligence. However, due to limitations in a lack of external feedback, these agents struggle with self-correction and generalization. A promising approach is to use reward models as external feedback, but there is no clear on how to select reward models for agents. Thus, there is an urgent need to build a reward bench targeted at agents. To address these challenges, we propose Agent-RewardBench, a benchmark designed to evaluate reward modeling ability in MLLMs. The benchmark is characterized by three key features: (1) Multiple dimensions and real-world agent scenarios evaluation. It covers perception, planning, and safety with 7 scenarios; (2) Step-level reward evaluation. It allows for the assessment of agent capabilities at the individual steps of a task, providing a more granular view of performance during the planning process; and (3) Appropriately difficulty and high-quality. We carefully sample from 10 diverse models, difficulty control to maintain task challenges, and manual verification to ensure the integrity of the data. Experiments demonstrate that even state-of-the-art multimodal models show limited performance, highlighting the need for specialized training in agent reward modeling. Code is available at github.

DNASpeech: A Contextualized and Situated Text-to-Speech Dataset with Dialogues, Narratives and Actions

Chuanqi Cheng, Hongda Sun, Bo Du, Shuo Shang, Xinrong Hu, Rui Yan

In this paper, we propose contextualized and situated text-to-speech (CS-TTS), a novel TTS task to promote more accurate and customized speech generation using prompts with Dialogues, Narratives, and Actions (DNA). While prompt-based TTS methods facilitate controllable speech generation, existing TTS datasets lack situated descriptive prompts aligned with speech data. To address this data scarcity, we develop an automatic annotation pipeline enabling multifaceted alignment among speech clips, content text, and their respective descriptions. Based on this pipeline, we present DNASpeech, a novel CS-TTS dataset with high-quality speeches with DNA prompt annotations. DNASpeech contains 2,395 distinct characters, 4,452 scenes, and 22,975 dialogue utterances, along with over 18 hours of high-quality speech recordings. To accommodate more specific task scenarios, we establish a leaderboard featuring two new subtasks for evaluation: CS-TTS with narratives and CS-TTS with dialogues. We also design an intuitive baseline model for comparison with existing state-of-the-art TTS methods on our leaderboard. Comprehensive experimental results demonstrate the quality and effectiveness of DNASpeech, validating its potential to drive advancements in the TTS field.

LLMs can be easily Confused by Instructional Distractions

Yerin Hwang, Yongil Kim, Jahyun Koo, Taegwan Kang, Hyunkyung Bae, Kyomin Jung

Despite the fact that large language models (LLMs) show exceptional skill in instruction following tasks, this strength can turn into a vulnerability when the models are required to disregard certain instructions. Instruction following tasks typically involve a clear task description and input text containing the target data to be processed. However, when the input itself resembles an instruction, confusion may arise, even if there is explicit prompting to distinguish between the task instruction and the input. We refer to this phenomenon as instructional distraction. In this paper, we introduce a novel benchmark, named **DIM-Bench**, specifically designed to assess LLMs' performance under instructional distraction. The benchmark categorizes real-world instances of instructional distraction and evaluates LLMs across four instruction tasks: proofreading, rewriting, translation, and style transfer---alongside five input tasks: reasoning, code generation, mathematical reasoning, bias detection, and question answering. Our experimental results reveal that even the most advanced LLMs are susceptible to instructional distraction, often failing to accurately follow user intent in such cases.

Dynamic Evaluation with Cognitive Reasoning for Multi-turn Safety of Large Language Models

Lanxue Zhang, Yanan Cao, Yuqiang Xie, Fang Fang, Yangxi Li

The rapid advancement of Large Language Models (LLMs) poses significant challenges for safety evaluation. Current static datasets struggle to identify emerging vulnerabilities due to three limitations: (1) they risk being exposed in model training data, leading to evaluation bias; (2) their limited prompt diversity fails to capture real-world application scenarios; (3) they are limited to provide human-like multi-turn interactions. To address these limitations, we propose a dynamic evaluation framework, CogSafe, for comprehensive and automated multi-turn safety assessment of LLMs. We introduce CogSafe based on cognitive theories to simulate the real chatting process. To enhance assessment diversity, we introduce scenario simulation and strategy decision to guide the dynamic generation, enabling coverage of application situations. Furthermore, we incorporate the cognitive process to simulate multi-turn dialogues that reflect the cognitive dynamics of real-world interactions. Extensive experiments demonstrate the scalability and effectiveness of our framework, which has been applied to evaluate the safety of widely used LLMs.

Unveiling Cultural Blind Spots: Analyzing the Limitations of mLLMs in Procedural Text Comprehension

Fajri Koto, Amir Hossein Yari

Despite the impressive performance of multilingual large language models (mLLMs) in various natural language processing tasks, their ability to understand procedural texts, particularly those with culture-specific content, remains largely unexplored. Texts describing cultural procedures, including rituals, traditional craftsmanship, and social etiquette, require an inherent understanding of cultural context, presenting a significant challenge for mLLMs. In this work, we introduce CAPTex, a benchmark designed to evaluate mLLMs' ability to process and reason over culturally diverse procedural texts in multiple languages. Using a range of evaluation methods, we find that (1) mLLMs struggle with culturally contextualized procedural content, particularly in low-resource languages; (2) performance varies across cultural domains, with some proving more difficult than others; and (3) models perform better on multiple-choice tasks presented in conversational formats than on direct questions. These results highlight the current limitations of mLLMs and emphasize the need for culturally informed benchmarks like CAPTex to support more accurate and inclusive language understanding.

Can LLMs Identify Critical Limitations within Scientific Research? A Systematic Evaluation on AI Research Papers

Zhijian Xu, Yilun Zhao, Manasi Patwardhan, Lovekesh Vig, Arman Cohan

Peer review is fundamental to scientific research, but the growing volume of publications has intensified the challenges of this expertise-intensive process. While LLMs show promise in various scientific tasks, their potential to assist with peer review, particularly in identifying paper limitations, remains understudied. We first present a comprehensive taxonomy of limitation types in scientific research, with a focus on AI. Guided by this taxonomy, for studying limitations, we present LimitGen, the first comprehensive benchmark for evaluating LLMs' capability to support early-stage feedback and complement human peer review. Our benchmark consists of two subsets: LimitGen-Syn, a synthetic dataset carefully created through controlled perturbations of high-quality papers, and LimitGen-Human, a collection of real human-written limitations. To improve the ability of LLM systems to identify limitations, we augment them with literature retrieval, which is essential for grounding identifying limitations in prior scientific findings. Our approach enhances the capabilities of LLM systems to generate limitations in research papers, enabling them to provide more concrete and constructive feedback.

CoIR: A Comprehensive Benchmark for Code Information Retrieval Models

Xiangyang Li, Kuicai Dong, Yi Quan Lee, Wei Xia, Hao Zhang, Xinyi Dai, Yasheng Wang, Ruiming Tang

Despite the substantial success of Information Retrieval (IR) in various NLP tasks, most IR systems predominantly handle queries and corpora in natural language, neglecting the domain of code retrieval. Code retrieval is critically important yet remains under-explored, with existing methods and benchmarks inadequately representing the diversity of code in various domains and tasks. Moreover, many models have begun to overfit existing leaderboards, limiting their generalizability and real-world applicability. Addressing this gap, we present CoIR (**Co**de **I**nformation **R**etrieval Benchmark), a robust and comprehensive benchmark specifically designed to assess code retrieval capabilities. CoIR comprises ten meticulously curated code datasets, spanning eight distinctive retrieval tasks across seven diverse domains. We first discuss the construction of CoIR and its diverse dataset composition. Further, we evaluate ten widely used retrieval models using CoIR, uncovering significant difficulties in performing code retrieval tasks even with state-of-the-art systems. CoIR also introduces a simple yet effective python framework, which additionally defines various advanced modes to facilitate researchers in evaluating their models. It shares the same data schema as other popular benchmarks like MTEB and BEIR, enabling seamless cross-benchmark evaluations. Through CoIR, we aim to invigorate research in the code retrieval domain, providing a versatile benchmarking tool that encourages further development and exploration of code retrieval systems.

MEMERAG: A Multilingual End-to-End Meta-Evaluation Benchmark for Retrieval Augmented Generation

María Andrea Cruz Blandón, Jayasimha Talur, Bruno Charron, Dong Liu, Saab Mansour, Marcello Federico

Automatic evaluation of retrieval augmented generation (RAG) systems relies on fine-grained dimensions like faithfulness and relevance, as judged by expert human annotators. Meta-evaluation benchmarks support the development of automatic evaluators that correlate well with human judgement. However, existing benchmarks predominantly focus on English or use translated data, which fails to capture cultural nuances. A native approach provides a better representation of the end user experience. In this work, we develop a Multilingual End-to-end Meta-Evaluation RAG benchmark MEMERAG. Our benchmark builds on the popular MIRACL dataset, using native-language questions and generating responses with diverse large language models (LLMs), which are then assessed by expert annotators for faithfulness and relevance. We describe our annotation process and show that it achieves high inter-annotator agreement. We then analyse the performance of the answer-generating LLMs across languages as per the human evaluators. Finally we apply the dataset to our main use-case which is to benchmark multilingual automatic evaluators (LLM-as-a-judge). We show that our benchmark can reliably identify improvements offered by advanced prompting techniques and LLMs. We release our benchmark to support the community developing accurate evaluation methods for multilingual RAG systems.

Around the World in 24 Hours: Probing LLM Knowledge of Time and Place

Carolin Holtermann, Paul Röttger, Anne Lauscher

Reasoning over time and space is essential for understanding our world. However, the abilities of language models in this area are largely unexplored as previous work has tested their abilities for logical reasoning in terms of time and space in isolation or only in simple or artificial environments. In this paper, we present the first evaluation of the ability of language models to jointly reason over time and space. To enable our analysis, we create GeoTemp, a dataset of 320k prompts covering 289 cities in 217 countries and 37 time zones. Using GeoTemp, we evaluate eight open chat models of three different model families for different combinations of temporal and geographic knowledge. We find that most models perform well on reasoning tasks involving only temporal knowledge and that overall performance improves with scale. However, performance remains constrained in tasks that require connecting temporal and geographical information. We do not find clear correlations of performance with specific geographic regions. Instead, we find a significant performance increase for location names with low model perplexity, suggesting their repeated occurrence during model training. We further demonstrate that their performance is heavily influenced by prompt formulation - a direct injection of geographical knowledge leads to performance gains, whereas, surprisingly, techniques like chain-of-thought prompting decrease performance on simpler tasks.

Ref-Long: Benchmarking the Long-context Referencing Capability of Long-context Language Models

Junjie Wu, Gefei Gu, Yanan Zheng, Dit-Yan Yeung, Arman Cohan

Long-context language models (LCLMs) have exhibited impressive capabilities in long-context understanding tasks. Among these, long-context referencing—a crucial task that requires LCLMs to attribute items of interest to specific parts of long-context data—remains underexplored. To bridge this gap, this paper proposes Referencing Evaluation for Long-context Language Models (Ref-Long), a novel benchmark designed to assess the long-context referencing capability of LCLMs. Specifically, Ref-Long requires LCLMs to identify the indexes of documents that reference a specific key, emphasizing contextual relationships between the key and the documents over simple retrieval. Based on the task design, we construct three subsets ranging from synthetic to realistic scenarios to form the Ref-Long benchmark. Experimental results of 13 LCLMs reveal significant shortcomings in long-context referencing, even among advanced models like GPT-4o. To further investigate these challenges, we conduct comprehensive analyses, including human evaluations, task format adjustments, fine-tuning experiments, and error analyses, leading to several key insights. Our data and code will be publicly released, and the data is also attached in the submission.

Revisiting Uncertainty Quantification Evaluation in Language Models: Spurious Interactions with Response Length Bias Results

Andrea Santilli, Adam Golinski, Michael Kirchhof, Federico Danieli, Arno Blaas, Miao Xiong, Luca Zappella, Sinead Williamson

Uncertainty Quantification (UQ) in Language Models (LMs) is key to improving their safety and reliability. Evaluations often use metrics like AUROC to assess how well UQ methods (e.g., negative sequence probabilities) correlate with task correctness functions (e.g., ROUGE-L). We show that mutual biases-when both UQ methods and correctness functions are biased by the same factors-systematically distort evaluation. First, we formally prove that any mutual bias non-randomly skews AUROC rankings, compromising benchmark integrity. Second, we confirm this happens empirically by testing 7 widely used correctness functions, from lexical-based and embedding-based metrics to LM-as-a-judge approaches, across 4 datasets × 4 models × 8 UQ methods. Our analysis showsthat length biases in correctness functions distort UQ assessments by interacting with length biases in UQ methods. We identify LM-as-a-judge methods as the least length-biased, offering a promising path for a fairer UQ evaluation.

HalluLens: LLM Hallucination Benchmark

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, Pascale Fung

Large language models (LLMs) often generate responses that deviate from user input or training data, a phenomenon known as "hallucination." These hallucinations undermine user trust and hinder the adoption of generative AI systems. Addressing hallucinations is important for the advancement of LLMs. This paper introduces a comprehensive hallucination benchmark HalluLens, incorporating both extrinsic and intrinsic evaluation tasks, built upon a clear taxonomy of hallucination. A major challenge in benchmarking hallucinations is the lack of a unified framework due to inconsistent definitions and categorizations. We disentangle LLM hallucination from "factuality" and propose a taxonomy distinguishing extrinsic and intrinsic hallucinations to promote consistency and facilitate research. We emphasize extrinsic hallucinations -- where generated content deviates from training data -- as they become increasingly relevant with LLM advancements. However, no benchmark is solely dedicated to extrinsic hallucinations. To address this gap, HalluLens introduces three new extrinsic tasks with dynamic test set generation to mitigate data leakage and ensure robustness. We release codebase for extrinsic hallucination benchmark.

Evaluating the Evaluation of Diversity in Commonsense Generation

Tianhui Zhang, Bei Peng, Danushka Bollegala

In commonsense generation, given a set of input concepts, a model must generate a response that is not only commonsense bearing, but also capturing multiple diverse viewpoints. Numerous evaluation metrics based on form- and content-level overlap have been proposed in prior work for evaluating the diversity of a commonsense generation model. However, it remains unclear as to which metrics are best suited for evaluating the diversity in commonsense generation. To address this gap, we conduct a systematic meta-evaluation of diversity metrics for commonsense generation. We find that form-based diversity metrics tend to consistently overestimate the diversity in sentence sets, where even randomly generated sentences are assigned overly high diversity scores. We then use an Large Language Model (LLM) to create a novel dataset annotated for the diversity of sentences generated for a commonsense generation task, and use it to conduct a meta-evaluation of the existing diversity evaluation metrics. Our experimental results show that content-based diversity evaluation metrics consistently outperform the form-based counterparts, showing high correlations with the LLM-based ratings. We recommend that future work on commonsense generation should use content-based metrics for evaluating the diversity of their outputs.

FastMCTS: A Simple Sampling Strategy for Data Synthesis

Peiji Li, Kai Lv, Yunfan Shao, Yichuan Ma, Linyang Li, Xiaoqing Zheng, Xipeng Qiu, Qipeng Guo

Synthetic high-quality multi-step reasoning data can significantly enhance the performance of large language models on various tasks. However, most existing methods rely on rejection sampling, which generates trajectories independently and suffers from inefficiency and imbalanced sampling across problems of varying difficulty. In this work, we introduce FastMCTS, an innovative data synthesis strategy inspired by Monte Carlo Tree Search. FastMCTS provides a more efficient sampling method for multi-step reasoning data, offering step-level evaluation signals and promoting balanced sampling across problems of different difficulty levels. Experiments on both English and Chinese reasoning datasets demonstrate that FastMCTS generates over 30% more correct reasoning paths compared to rejection sampling as the number of generated tokens scales up. Furthermore, under comparable synthetic data budgets, models trained on FastMCTS-generated data outperform those trained on rejection sampling data by 3.9% across multiple benchmarks. As a lightweight sampling strategy, FastMCTS offers a practical and efficient alternative for synthesizing high-quality reasoning data.

ACORD: An Expert-Annotated Retrieval Dataset for Legal Contract Drafting

Steven H Wang, Maksim Zubkov, Kexin Fan, Sarah Harrell, Yuyang Sun, Wei Chen, Andreas Plesner, Roger Wattenhofer

Contract clause retrieval is foundational to contract drafting because lawyers rarely draft contracts from scratch; instead, they locate and revise the most relevant precedent clauses. We introduce the Atticus Clause Retrieval Dataset (ACORD), the first expert-annotated benchmark specifically designed for contract clause retrieval to support contract drafting tasks. ACORD focuses on complex contract clauses such as Limitation of Liability, Indemnification, Change of Control, and Most Favored Nation. It includes 114 queries and over 126,000 query-clause pairs, each ranked on a scale from 1 to 5 stars. The task is to find the most relevant precedent clauses to a query. The bi-encoder retriever paired with pointwise LLMs re-rankers shows promising results. However, substantial improvements are still needed to manage the complex legal work typically undertaken by lawyers effectively. As the first expert-annotated benchmark for contract clause retrieval, ACORD can serve as a valuable IR benchmark for the NLP community.

The Nature of NLP: Analyzing Contributions in NLP Papers

Aniket Pramanick, Yufang Hou, Saif M. Mohammad, Iryna Gurevych

Natural Language Processing (NLP) is an established and dynamic field. Despite this, what constitutes NLP research remains debated. In this work, we address the question by quantitatively examining NLP research papers. We propose a taxonomy of research contributions and introduce _NLPContributions_, a dataset of nearly 2k NLP research paper abstracts, carefully annotated to identify scientific contributions and classify their types according to this taxonomy. We also introduce a novel task of automatically identifying contribution statements and classifying their types from research papers. We present experimental results for this task and apply our model to \sim29k NLP research papers to analyze their contributions, aiding in the understanding of the nature of NLP research. We show that NLP research has taken a winding path --- with the focus on language and human-centric studies being prominent in the 1970s and 80s, tapering off in the 1990s and 2000s, and starting to rise again since the late 2010s. Alongside this revival, we observe a steady rise in dataset and methodological contributions since the 1990s, such that today, on average, individual NLP papers contribute in more ways than ever before. Our dataset and analyses offer a powerful lens for tracing research trends and offer potential for generating informed, data-driven literature surveys.

ASPERA: A Simulated Environment to Evaluate Planning for Complex Action Execution

Alexandru Coca, Mark Gaynor, Zhenxing Zhang, Jianpeng Cheng, Bo-Hsiang Tseng, Peter Boothroyd, Hector Martinez Alonso, Diarmuid O Seaghdha, Anders Johannsen

This work evaluates the potential of large language models (LLMs) to power digital assistants capable of complex action execution. Such assistants rely on pre-trained programming knowledge to execute multi-step goals by composing objects and functions defined in assistant libraries into action execution programs. To achieve this, we develop ASPERA, a framework comprising an assistant library simulation and a human-assisted LLM data generation engine. Our engine allows developers to guide LLM generation of high-quality tasks consisting of complex user queries, simulation state and corresponding validation programs, tackling data availability and evaluation robustness challenges. Alongside the framework we release Asper-Bench, an evaluation dataset of 250 challenging tasks generated using ASPERA, which we use to show that program generation grounded in custom assistant libraries is a significant challenge to LLMs compared to dependency-free code generation.

Improving Automatic Evaluation of Large Language Models (LLMs) in Biomedical Relation Extraction via LLMs-as-the-Judge

Md Tahmid Rahman Laskar, Israt Jahan, Elham Dolatabadi, Chun Peng, Enamul Hoque, Jimmy Huang

Large Language Models (LLMs) have demonstrated impressive performance in biomedical relation extraction, even in zero-shot scenarios. However, evaluating LLMs in this task remains challenging due to their ability to generate human-like text, often producing synonyms or abbreviations of gold-standard answers, making traditional automatic evaluation metrics unreliable. On the other hand, while human evaluation is more reliable, it is costly and time-consuming, making it impractical for real-world applications. This paper investigates the use of LLMs-as-the-Judge as an alternative evaluation method for biomedical relation extraction. We benchmark 8 LLMs as judges to evaluate the responses generated by 5 other LLMs across 3 biomedical relation extraction datasets. Unlike other text-generation tasks, we observe that LLM-based judges perform quite poorly (usually below 50% accuracy) in the biomedical relation extraction task. Our findings reveal that it happens mainly because relations extracted by LLMs do not adhere to any standard format. To address this, we propose structured output formatting for LLM-generated responses that helps LLM-Judges to improve their performance by about 15% (on average). We also introduce a domain adaptation technique to further enhance LLM-Judge performance by effectively transferring knowledge between datasets. We release both our human-annotated and LLM-annotated judgment data (36k samples in total) for public use here: https://github.com/tahmedge/llm_judge_biomedical_re.

LogicPro: Improving Complex Logical Reasoning via Program-Guided Learning

Jin Jiang, Yuchen Yan, Yang Liu, Jianing Wang, Shuai Peng, Xunliang Cai, Yixin Cao, Mengdi Zhang, Liangcai Gao

In this paper, we propose a new data synthesis method called \textbf{LogicPro}, which leverages LeetCode-style algorithm \underline{Pro}blems and their corresponding \underline{Pro}gram solutions to synthesize Complex \underline{Logic}al Reasoning data in text format. First, we synthesize complex reasoning problems through source algorithm problems and test cases. Then, standard answers and intermediate variable outputs are obtained for each problem based on standard python solutions and test cases. Finally, with the guidance of code intermediate variables, we synthesize the text reasoning process for each reasoning problems. Through this method, we can synthesize data that is difficult, scalable, effective, and comes with golden standard answers and high-quality reasoning processes. As a result, with our 540K synthesized dataset constructed solely from 2,360 algorithm problems, our approach achieves significant improvements in multiple models for the datasets \textit{BBH^27}, \textit{LogicBench}, \textit{DROP}, \textit{AR-LSAT}, and \textit{GSM8K}, etc. outperforming a wide range of existing reasoning datasets.

InductionBench: LLMs Fail in the Simplest Complexity Class

Wenyue Hua, Tyler Wong, Fei Sun, Liangming Pan, Adam Jardine, William Yang Wang

Large language models (LLMs) have shown remarkable improvements in reasoning and many existing benchmarks have been addressed by models such as o1 and o3 either fully or partially. However, a majority of these benchmarks emphasize deductive reasoning, including mathematical and coding tasks in which rules such as mathematical axioms or programming syntax are clearly defined, based on which LLMs can plan and apply these rules to arrive at a solution. In contrast, \textit{inductive reasoning}, where one infers the underlying rules from observed data, remains less explored. Such inductive processes lie at the heart of scientific discovery, as they enable researchers to extract general principles from empirical observations. To assess whether LLMs possess this capacity, we introduce \textbf{InductionBench}, a new benchmark designed to evaluate the inductive reasoning ability of LLMs. Our experimental findings reveal that even the most advanced modelw available struggle to master the simplest complexity classes within the subregular hierarchy of functions, highlighting a notable deficiency in current LLMs' inductive reasoning capabilities. Coda and data are available .

From Citations to Criticality: Predicting Legal Decision Influence in the Multilingual Swiss Jurisprudence

Ronja Stern, Ken Kawamura, Matthias Stürmer, Ilias Chalkidis, Joel Niklaus

Many court systems are overwhelmed all over the world, leading to huge backlogs of pending cases. Effective triage systems, like those in emergency rooms, could ensure proper prioritization of open cases, optimizing time and resource allocation in the court system. In this work, we introduce the Criticality Prediction dataset, a novel resource for evaluating case prioritization. Our dataset features a two-tier labeling system: (1) the binary LD-Label, identifying cases published as Leading Decisions (LD), and (2) the more granular Citation-Label, ranking cases by their citation frequency and recency, allowing for a more nuanced evaluation. Unlike existing approaches that rely on resource-intensive manual annotations, we algorithmically derive labels leading to a much larger dataset than otherwise possible. We evaluate several multilingual models, including both smaller fine-tuned models and large language models in a zero-shot setting. Our results show that the fine-tuned models consistently outperform their larger counterparts, thanks to our large training set. Our results highlight that for highly domain-specific tasks like ours, large training sets are still valuable.

A Dual-Perspective NLG Meta-Evaluation Framework with Automatic Benchmark and Better Interpretability

Xinyu Hu, Mingqi Gao, Li Lin, Zhenghan Yu, Xiaojun Wan

In NLG meta-evaluation, evaluation metrics are typically assessed based on their consistency with humans. However, we identify some limitations in traditional NLG meta-evaluation approaches, such as issues in handling human ratings and ambiguous selections of correlation measures, which undermine the effectiveness of meta-evaluation. In this work, we propose a dual-perspective NLG meta-evaluation framework that focuses on different evaluation capabilities, thereby providing better interpretability. In addition, we introduce a method of automatically constructing the corresponding benchmarks without requiring new human annotations. Furthermore, we conduct experiments with 16 representative LLMs as the evaluators based on our proposed framework, comprehensively analyzing their evaluation performance from different perspectives.

Evaluating Sequence Labeling on the basis of Information Theory

Enrique Amigo, Elena Álvarez-Mellado, Julio Gonzalo, Jorge Carrillo-de-Albornoz

Various metrics exist for evaluating sequence labeling problems (strict span matching, token oriented metrics, token concurrence in sequences, etc.), each of them focusing on certain aspects of the task. In this paper, we define a comprehensive set of formal properties that captures the strengths and weaknesses of the existing metric families and prove that none of them is able to satisfy all properties simultaneously. We argue that it is necessary to measure how much information (correct or noisy) each token in the sequence contributes depending on different aspects such as sequence length, number of tokens annotated by the system, token specificity, etc. On this basis, we introduce the {\bf S}equence {\bf L}abelling {\bf I}nformation {\bf C}ontrast {\bf M}odel (SL-ICM), a novel metric based on information theory for evaluating sequence labeling tasks. Our formal analysis and experimentation show that the proposed metric satisfies all properties simultaneously

AgentGym: Evaluating and Training Large Language Model-based Agents across Diverse Environments

Zhiheng Xi, Yiwen Ding, Wenxiang Chen, Boyang Hong, Honglin Guo, Junzhe Wang, Xin Guo, Dingwen Yang, Chenyang Liao, Wei He, Songyang Gao, Lu Chen, Rui Zheng, Yicheng Zou, Tao Gui, Qi Zhang, Xipeng Qiu, Xuanjing Huang, Zuxuan Wu, Yu-Gang Jiang

Large language models (LLMs) have emerged as a promising foundation to build generally-capable agents (LLM-based agents) that can handle multi-turn decision-making tasks across various environments. However, the community lacks a unified interactive framework that covers diverse environments for comprehensive evaluation of agents, and enables exploration and learning for their self-improvement. To address this, we propose AgentGym, a framework featuring 7 real-world scenarios, 14 environments, and 89 tasks for unified, real-time, and concurrent agent interaction. We construct expanded instruction set, high-quality trajectories, and comprehensive benchmarking suite for developing LLM-based agents. Moreover, AgentGym supports interactive exploration and learning for agents through multi-turn interactions and real-time feedback. Based on AgentGym, we take the initial step to develop LLM-based agents that can handle diverse tasks via methods like self-improvement or reinforcement learning. Experimental results show that the trained agents can achieve results comparable to commercial models. We hope our work can help the community develop more advanced LLM-based agents. We release the code, dataset, benchmark, and checkpoints at https://agentgym.github.io/.

Adverse Event Extraction from Discharge Summaries: A New Dataset, Annotation Scheme, and Initial Findings

Imane Guellil, Salomé Andres, Atul Anand, Bruce Guthrie, Huayu Zhang, Abul Hasan, Honghan Wu, Beatrice Alex

In this work, we present a manually annotated corpus for Adverse Event (AE) extraction from discharge summaries of elderly patients, a population often underrepresented in clinical NLP resources. The dataset includes 14 clinically significant AEs—such as falls, delirium, and intracranial haemorrhage, along with contextual attributes like negation, diagnosis type, and in-hospital occurrence. Uniquely, the annotation schema supports both discontinuous and overlapping entities, addressing challenges rarely tackled in prior work. We evaluate multiple models using FlairNLP across three annotation granularities: fine-grained, coarse-grained, and coarse-grained with negation. While transformer-based models (e.g., BERT-cased) achieve strong performance on document-level coarse-grained extraction (F1 = 0.943), performance drops notably for fine-grained entity-level tasks (e.g., F1 = 0.675), particularly for rare events and complex attributes. These results demonstrate that despite high-level scores, significant challenges remain in detecting underrepresented AEs and capturing nuanced clinical language. Developed within a Trusted Research Environment (TRE), the dataset is available upon request via DataLoch and serves as a robust benchmark for evaluating AE extraction methods and supporting future cross-dataset generalisation.

From Neurons to Semantics: Evaluating Cross-Linguistic Alignment Capabilities of Large Language Models via Neurons Alignment

Chongxuan Huang, Yongshi Ye, Biao Fu, Qifeng Su, Xiaodong Shi

Large language models (LLMs) have demonstrated remarkable multilingual capabilities, however, how to evaluate cross-lingual alignment remains underexplored. Existing alignment benchmarks primarily focus on sentence embeddings, but prior research has shown that neural models tend to induce a non-smooth representation space, which impact of semantic alignment evaluation on low-resource languages. Inspired by neuroscientific findings that similar information activates overlapping neuronal regions, we propose a novel *Neuron State-Based Cross-Lingual Alignment* (\textbf{NeuronXA}) to assess the cross-lingual a lignment capabilities of LLMs, which offers a more semantically grounded approach to assess cross-lingual alignment. We evaluate NeuronXA on several prominent multilingual LLMs (LLaMA, Qwen, Mistral, GLM, and OLMo) across two transfer tasks and three multilingual benchmarks. The results demonstrate that with only 100 parallel sentence pairs, NeuronXA achieves a Pearson correlation of 0.9556 with downstream tasks performance and 0.8524 with transferability. These findings demonstrate NeuronXA's effectiveness in assessing both cross-lingual alignment and transferability, even with a small dataset. This highlights its potential to advance cross-lingual alignment research and to improve the semantic understanding of multilingual LLMs.

Cross-Lingual Auto Evaluation for Assessing Multilingual LLMs

Sumanth Doddapaneni, Mohammed Safi Ur Rahman Khan, Dilip Venkatesh, Raj Dabre, Anoop Kunchukuttan, Mitesh M Khapra

Evaluating machine-generated text remains a significant challenge in NLP, especially for non-English languages. Current methodologies, including automated metrics, human assessments, and LLM-based evaluations, predominantly focus on English, revealing a significant gap in multilingual evaluation frameworks. We introduce the Cross Lingual Auto Evaluation (CIA) Suite, an extensible framework that includes evaluator LLMs (Hercule) and a novel test set (Recon) specifically designed for multilingual evaluation. Our test set features 500 human-annotated instructions spanning various task capabilities along with human judgment scores across six languages. This would enable benchmarking of general-purpose multilingual LLMs and facilitate meta-evaluation of Evaluator LLMs. The proposed model, Hercule, is a cross-lingual evaluation model that addresses the scarcity of reference answers in the target language by learning to assign scores to responses based on easily available reference answers in English. Our experiments demonstrate that Hercule aligns more closely with human judgments compared to proprietary models, demonstrating the effectiveness of such cross-lingual evaluation in low resource scenarios. Further, it is also effective in zero-shot evaluation on unseen languages. This study is the first comprehensive examination of cross-lingual evaluation using LLMs, presenting a scalable and effective approach for multilingual assessment. All code, datasets, and models will be publicly available to enable further research in this important area.

Zero-Shot Text-to-Speech for Vietnamese

Thi Vu, Linh The Nguyen, Dat Quoc Nguyen

This paper introduces PhoAudiobook, a newly curated dataset comprising 941 hours of high-quality audio for Vietnamese text-to-speech. Using PhoAudiobook, we conduct experiments on three leading zero-shot TTS models: VALL-E, VoiceCraft, and XTTS-V2. Our findings demonstrate that PhoAudiobook consistently enhances model performance across various metrics. Moreover, VALL-E and VoiceCraft exhibit superior performance in synthesizing short sentences, highlighting their robustness in handling diverse linguistic contexts. We publicly release PhoAudiobook to facilitate further research and development in Vietnamese text-to-speech.

Batayan: A Filipino NLP benchmark for evaluating Large Language Models

Jann Railey Montalan, Jimson Paulo Layacan, David Demitri Africa, Richell Isaiah S. Flores, Michael T. Lopez II, Theresa Denise Magsajo, Anjanette Cayabyab, William Chandra Tjhi

Recent advances in large language models (LLMs) have demonstrated remarkable capabilities on widely benchmarked high-resource languages. However, linguistic nuances of under-resourced languages remain unexplored. We introduce Batayan, a holistic Filipino benchmark that systematically evaluates LLMs across three key natural language processing (NLP) competencies: understanding, reasoning, and generation. Batayan consolidates eight tasks, three of which have not existed prior for Filipino corpora, covering both Tagalog and code-switched Taglish utterances. Our rigorous, native-speaker-driven adaptation and validation processes ensures fluency and authenticity to the complex morphological and syntactic structures of Filipino, alleviating the pervasive translationese bias in existing Filipino corpora. We report empirical results on a variety of open-source and commercial LLMs, highlighting significant performance gaps that signal the under-representation of Filipino in pre-training corpora, the unique hurdles in modeling Filipino's rich morphology and construction, and the importance of explicit Filipino language support. Moreover, we discuss the practical challenges encountered in dataset construction and propose principled solutions for building culturally and linguistically-faithful resources in under-represented languages. We also provide a public evaluation suite as a clear foundation for iterative, community-driven progress in Filipino NLP.

SeedBench: A Multi-task Benchmark for Evaluating Large Language Models in Seed Science

Jie Ying, Zihong Chen, Zhefan Wang, Wanli Jiang, Chenyang Wang, Zhonghang Yuan, Haoyang Su, Huanjun Kong, Fan Yang, Nanqing Dong

Seed science is essential for modern agriculture, directly influencing crop yields and global food security. However, challenges such as interdisciplinary complexity and high costs with limited returns hinder progress, leading to a shortage of experts and insufficient technological support. While large language models (LLMs) have shown promise across various fields, their application in seed science remains limited due to the scarcity of digital resources, complex gene-trait relationships, and the lack of standardized benchmarks. To address this gap, we introduce SeedBench—the first multi-task benchmark specifically designed for seed science. Developed in collaboration with domain experts, SeedBench focuses on seed breeding and simulates key aspects of modern breeding processes. We conduct a comprehensive evaluation of 26 leading LLMs, encompassing proprietary, open-source, and domain-specific fine-tuned models. Our findings not only highlight the substantial gaps between the power of LLMs and the real-world seed science problems, but also make a foundational step for research on LLMs for seed design.

WiCkeD: A Simple Method to Make Multiple Choice Benchmarks More Challenging

Ahmed Elhady, Eneko Agirre, Mikel Artetxe

We introduce WiCkeD, a simple method to increase the complexity of existing multiple-choice benchmarks by randomly replacing a choice with "None of the above", a method often used in educational tests. We show that WiCkeD can be automatically applied to any existing benchmark, making it more challenging. We apply WiCkeD to 6 popular benchmarks and use it to evaluate 18 open-weight LLMs. The performance of the models drops12.1 points on average with respect to the original versions of the datasets. When using chainof-thought on 3 MMLU datasets, the performance drop for the WiCkeD variant is similar to the one observed when using the LLMs directly, showing that WiCkeD is also challenging for models with enhanced reasoning abilities. WiCkeD also uncovers that some models are more sensitive to the extra reasoning required, providing additional information with respect to the original benchmarks.We relase our code and data at github.com/anonymized.

ONEBench to Test Them All: Sample-Level Benchmarking Over Open-Ended Capabilities

Adhiraj Ghosh, Sebastian Dziadzio, Ameya Prabhu, Vishaal Udandarao, Samuel Albanie, Matthias Bethge

Traditional fixed test datasets fall short in evaluating the open-ended capabilities of foundation models. To address this, we propose ONEBench (OpeN-Ended Benchmarking), a new paradigm that consolidates individual evaluation datasets into a unified, ever-expanding sample pool. ONEBench enables custom benchmarks for specific capabilities while reusing and aggregating samples, mitigating overfitting and dataset bias for broader capability assessment. It reframes model evaluation as selecting and aggregating sample-level tests. Transitioning from task-specific benchmarks to ONEBench introduces two challenges: heterogeneity (aggregating diverse metrics) and incompleteness(comparing models tested on different data subsets). To address these, we propose an aggregation algorithm that ensures identifiability (asymptotically recovering ground-truth scores) and rapid convergence, enabling accurate model comparisons with relatively little data. On homogenous datasets, our algorithm produces rankings that highly correlate with average scores. Moreover, it remains robust to over 95% missing measurements, reducing evaluation costs by up to 20x with minimal impact on rankings. We introduce ONEBench-LLM for language models and ONEBench-LMM for vision-language models, unifying evaluations across these domains, and enabling targeted model testing across diverse capabilities.

La Leaderboard: A Large Language Model Leaderboard for Spanish Varieties and Languages of Spain and Latin America

María Grandury, Javier Aula-Blasco, Júlia Falcão, Clémentine Fourrier, Miguel González Saiz, Gonzalo Martínez, Gonzalo Santamaria Gomez, Rodrigo Agerri, Nuria Aldama García, Luis Chiruzzo, Javier Conde, Helena Gomez Adorno, Marta Guerrero Nieto, Guido Ivetta, Natàlia López Fuertes, Flor Miriam Plaza-del-Arco, María-Teresa Martín-Valdivia, Helena Montoro Zamorano, Carmen Muñoz Sanz, Pedro Reviriego, Leire Rosado Plaza, Alejandro Vaca Serrano, Estrella Vallecillo-Rodríguez, Jorge Vallego, Irune Zubiaga

Leaderboards showcase the current capabilities and limitations of Large Language Models (LLMs). To motivate the development of LLMs that represent the linguistic and cultural diversity of the Spanish-speaking community, we present La Leaderboard, the first open-source leaderboard to evaluate generative LLMs in languages and language varieties of Spain and Latin America. La Leaderboard is a community-driven project that aims to establish an evaluation standard for everyone interested in developing LLMs for the Spanish-speaking community. This initial version combines 66 datasets in Catalan, Basque, Galician, and different Spanish varieties, showcasing the evaluation results of 50 models. To encourage community-driven development of leaderboards in other languages, we explain our methodology, including guidance on selecting the most suitable evaluation setup for each downstream task. In particular, we provide a rationale for using fewer few-shot examples than typically found in the literature, aiming to reduce environmental impact and facilitate access to reproducible results for a broader research community.

LangMark: A Multilingual Dataset for Automatic Post-Editing

Diego Velazquez, Mikaela Grace, Konstantinos Karageorgos, Lawrence Carin, Aaron Schliem, Dimitrios Zaikis, Roger Wechsler

Automatic post-editing (APE) aims to correct errors in machine-translated text, enhancing translation quality, while reducing the need for human intervention. Despite advances in neural machine translation (NMT), the development of effective APE systems has been hindered by the lack of large-scale multilingual datasets specifically tailored to NMT outputs. To address this gap, we present and release LangMark, a new human-annotated multilingual APE dataset for English translation to seven languages: Brazilian Portuguese, French, German, Italian, Japanese, Russian, and Spanish. The dataset has 206,983 triplets, with each triplet consisting of a source segment, its NMT output, and a human post-edited translation. Annotated by expert human linguists, our dataset offers both linguistic diversity and scale. Leveraging this dataset, we empirically show that Large Language Models (LLMs) with few-shot prompting can effectively perform APE, improving upon leading commercial and even proprietary machine translation systems. We believe that this new resource will facilitate the future development and evaluation of APE systems.

KRISTEVA: Close Reading as a Novel Task for Benchmarking Interpretive Reasoning

Peiqi Sui, Juan Diego Rodriguez, Philippe Laban, J. Dean Murphy, Joseph P. Dexter, Richard Jean So, Samuel Baker, Pramit Chaudhuri

Each year, tens of millions of essays are written and graded in college-level English courses. Students are asked to analyze literary and cultural texts through a process known as close reading, where they gather textual details from which to formulate evidence-based arguments. Despite being viewed as a basis for critical thinking and widely adopted as a required element of university coursework, close reading has never been evaluated on large language models (LLMs), and multi-discipline benchmarks like MMLU do not include literature as a subject. To fill this gap, we present KRISTEVA, the first close reading benchmark for evaluating interpretive reasoning, consisting of 1331 multiple-choice questions adapted from classroom data. With KRISTEVA, we propose three progressively more difficult sets of tasks to approximate different elements of the close reading process, which we use to test how well LLMs understand and reason about literary works: 1) extracting stylistic features, 2) retrieving relevant contextual information from parametric knowledge, and 3) multi-hop reasoning between style and external contexts. Our baseline results find that while state-of-the-art LLMs possess some college-level close reading competency (accuracy 49.7% - 69.7%), their performances still trail those of experienced human evaluators on 10 out of our 11 tasks.

Session 12: Sentiment Analysis, Stylistic Analysis, and Argument Mining (Wednesday 30 July, 11:00-12:30, Hall 5X)

The Open Argument Mining Framework

Debela Gemechu, Ramon Ruiz-Dolz, Somaye Moslemnejad, Eimear Maguire, Yohan Jo, John Lawrence, Chris Reed

Despite extensive research in Argument Mining (AM), the field faces significant challenges in limited reproducibility, difficulty in comparing systems due to varying task combinations, and a lack of interoperability caused by the heterogeneous nature of argumentation theory. These challenges are further exacerbated by the absence of dedicated tools, with most advancements remaining isolated research outputs rather than reusable systems. The \texttt{oAMF} (Open Argument Mining Framework) addresses these issues by providing an open-source, modular, and scalable platform that unifies diverse AM methods. Initially released with seventeen integrated modules, the \texttt{oAMF} serves as a starting point for researchers and developers to build, experiment with, and deploy AM pipelines while ensuring interoperability and allowing multiple theories of argumentation to co-exist within the same framework. Its flexible design supports integration via Python APIs, drag-and-drop tools, and web interfaces, streamlining AM development for research and industry setup, facilitating method comparison, and reproducibility.

DISPUTool 3.0: Fallacy Detection and Repairing in Argumentative Political Debates

Pierpaolo Goffredo, Deborah Dore, Elena Cabrio, Serena Villata

This paper introduces and evaluates a novel web-based application designed to identify and repair fallacious arguments in political debates. DISPUTool 3.0 offers a comprehensive tool for argumentation analysis of political debate, integrating state-of-the-art natural language processing techniques to mine and classify argument components and relations. DISPUTool 3.0 builds on the \textit{ElecDeb60to20} dataset, covering US presidential debates from 1960 to 2020. In this paper, we introduce a novel task which is integrated as a new module in DISPUTool, i.e., the automatic detection and classification of fallacious arguments, and the automatic \textit{repairing} of such misleading arguments. The goal is to show to the user a tool which not only identifies fallacies in political debates, but it also shows how the argument looks like once the veil of fallacy falls down. An extensive evaluation of the module is addressed employing both automated metrics and human assessments. With the inclusion of this module, DISPUTool 3.0 advances even more user critical thinking in front of the augmenting spread of such nefarious kind of content in political debates and beyond. The tool is publicly available here: https://3ia-demos.inria.fr/disputool/

Session 12: TACL Journal (Wednesday 30 July, 11:00-12:30, Hall 4/5)

A Comparative Approach for Auditing Multilingual Phonetic Transcript Archives

Farhan Samir, Emily, Shreya Prakash, Márton Sóskuthy, Vered Shwartz, Jian Zhu

Curating datasets that span multiple languages is challenging. To make the collection more scalable, researchers often incorporate one or more imperfect classifiers in the process, like language identification models. These models, however, are prone to failure, resulting in some language partitions being unreliable for downstream tasks. We introduce a statistical test, the Preference Proportion Test, for identifying such unreliable partitions. By annotating only 20 samples for a language partition, we're able to identify systematic transcription errors for 10 language partitions in a recent large multilingual transcribed audio archive, X-IPAPACK (Zhu et al., 2024). We find that filtering these low-quality partitions out when training models for the downstream task of phonetic transcription brings substantial benefits, most notably a 25.7% relative improvement on transcribing recordings in out-of-distribution languages. Our work contributes an effective method for auditing multilingual audio archives.

TANQ: An open domain dataset of table answered questions

Mubashara Akhtar, Chenxi Pang, Andreea Marzoca, Yasemin Altun, Julian Eisenschlos

Language models, potentially augmented with tool usage such as retrieval are becoming the go-to means of answering questions. Understanding and answering questions in real-world settings often requires re-trieving information from different sources, processing and aggregating data to extract insights, and presenting complex findings in form of structured artifacts such as novel tables, charts, or infographics. In this paper, we introduce TANQ, the first open domain question answering dataset where the answers require building tables from information across multiple sources. We release the full source attribution for every cell in the resulting table and benchmark state-of-the-art language models in open, oracle, and closed book setups. Our best-performing baseline, Gemini Flash reaches an overall F1 score of 60.7, lagging behind human performance by 12.3 points. We analyse baselines' performance across different dataset attributes such as different skills required for this task, including multi-hop reasoning, math operations, and unit conversions. We further discuss common failures in model-generated answers, suggesting that TANQ is a complex task with many challenges ahead.

STPar:AStructure-AwareTriaffineParserforScreenplay CharacterCoreferenceResolution

Li Zheng, Hao Fei, Lei Chen, Bobo Li, Fei Li, Chong Teng, Liang Zhao, Donghong Ji

Character Coreference Resolution in Movie Screenplays (MovieCoref) is a newly emerging task for understanding complex movie plots and character relationships. This task poses greater challenges than traditional coreference resolution, due to the intricate narrative structures and character interactions unique to screenplays. In light of these challenges, we introduce a novel approach: a Structure-aware Triaffine Parser (STPar) for the MovieCoref task. The STPar combines discourse and syntactic structures in the feature encoding process, enabling comprehensive analysis of ternary relationships and complex interactions. During the pairing process, the STPar utilizes a triaffine scorer to consider high-order relations between candidate mention pairs, thus enhancing its ability to capture detailed narrative structures. In addition, the STPar incorporates multi-task learning, encompassing singleton and span detection tasks, to further improve coreference resolution performance. Our evaluations on the MovieCoref dataset demonstrate that the STPar significantly outperforms the best baseline by 6.9\%, 21.1\%, 6.7\% and 9.7\% in F1 scores of B^3, CEAF_e, LEA and CoNLL. Further analysis highlights the benefits of integrating structural discourse and syntactic information as well as the combined approaches of triaffine and multi-task learning.

Sense-Specific Historical Word Usage Generation

Pierluigi Cassotti, Nina Tahmasebi

Large-scale sense-annotated corpora are important for a range of tasks but are hard to come by. Dictionaries that record and describe the vocabulary of a language often offer a small set of real-world example sentences for each sense of a word. However, on their own, these sentences are too few to be used as diachronic sense-annotated corpora. We propose a targeted strategy for training and evaluating generative models producing historically and semantically accurate word usages given any word, sense definition, and year triple. Our results demonstrate that the fine-tuned models can generate usages with the same properties as real-world example sentences from a reference dictionary. Thus the generated usages will be suitable for training and testing computational models where large-scale sense-annotated corpora are needed but currently unavailable.

Navigating the Landscape of Hint Generation Research: From the Past to the Future

Anubhav Jangra, Jamshid Mozafari, Adam Jatowt, Smaranda Muresan

Digital education has gained popularity in the last decade, especially after the COVID-19 pandemic. With the improving capabilities of large language models to reason and communicate with users, envisioning intelligent tutoring systems (ITSs) that can facilitate self-learning is not very far-fetched. One integral component to fulfill this vision is the ability to give accurate and effective feedback via hints to scaffold the learning process. In this survey article, we present a comprehensive review of prior research on hint generation, aiming to bridge the gap between research in education and cognitive science, and research in AI and Natural Language Processing. Informed by our findings, we propose a formal definition of the hint generation task, and discuss the roadmap of building an effective hint generation system aligned with the formal definition, including open challenges, future directions and ethical considerations.

MTRAG: A Multi-Turn Conversational Benchmark for Evaluating Retrieval-Augmented Generation Systems

Yannis Katsis, Sara Rosenthal, Kshitij Fadnis, Chulaka Gunasekara, Young-Suk Lee, Lucian Popa, Vraj Shah, Huaiyu Zhu, Marina Danilevsky

Retrieval-augmented generation (RAG) has recently become a very popular task for Large Language Models (LLMs). Evaluating them on multi-turn RAG conversations, where the system is asked to generate a response to a question in the context of a preceding conversation is an important and often overlooked task with several additional challenges. We present MTRAG: an end-to-end human-generated multi-turn RAG benchmark that reflects several real-world properties across diverse dimensions for evaluating the full RAG pipeline. MTRAG contains 110 conversations averaging 7.7 turns each across four domains for a total of 842 tasks. We also explore automation paths via synthetic data and LLM-as-a-Judge evaluation. Our human and automatic evaluations show that even state-of-the-art LLM RAG systems struggle on MTRAG. We demonstrate the need for strong retrieval and generation systems that can handle later turns, unanswerable questions, non-standalone questions, and multiple domains. MTRAG is available at https://github.com/ibm/mt-rag-benchmark.

Exploring Practical Gaps in Using Cross Entropy to Implement Maximum Mutual Information Criterion for Rationalization

Wei Liu, Zhiying Deng, Zhongyu Niu, Jun Wang, Haozhao Wang, Ruixuan Li

Rationalization is a framework that aims to build self-explanatory NLP models by extracting a subset of human-intelligible pieces of their inputting texts. It involves a cooperative game where a selector selects the most human-intelligible parts of the input as the rationale, followed by a predictor that makes predictions based on these selected rationales. Existing literature uses the cross-entropy between the model's predictions and the ground-truth labels to measure the informativeness of the selected rationales, guiding the selector to choose better ones. In this study, we first theoretically analyze the objective of rationalization by decomposing it into two parts: the model-agnostic informativeness of the rationale candidates and the predictor's degree of fit. We then provide various empirical evidence to support that, under this framework, the selector tends to sample from a limited small region, causing the predictor to overfit these localized areas. This results in a significant mismatch between the cross-entropy objective and the informativeness of the rationale candidates, leading to suboptimal solutions. To address this issue, we propose a simple yet effective method that introduces random vicinal perturbations to the selected rationale candidates. This approach broadens the predictor's assessment to a vicinity around the selected rationale candidate. Compared to recent competitive methods, our method significantly improves rationale quality (by up to 6.6\%) across six widely used classification datasets. Further experiments show that it can also generalize to the reading comprehension task and the fact extraction and verification task.

Human Choice Prediction in Language-Based Persuasion Games: Simulation-Based Off-Policy Evaluation

Eilam Shapira, Omer Madmon, Reut Apel, Moshe Tennenholtz, Roi Reichart

Recent advances in Large Language Models (LLMs) have spurred interest in designing LLM-based agents for tasks that involve interaction with human and artificial agents. This paper addresses a key aspect in the design of such agents: predicting human decisions in off-policy evaluation (OPE). We focus on language-based persuasion games, where an expert aims to influence the decision-maker through verbal messages. In our OPE framework, the prediction model is trained on human interaction data collected from encounters with one set of expert agents, and its performance is evaluated on interactions with a different set of experts. Using a dedicated application, we collected a dataset of 87K decisions from humans playing a repeated decision-making game with artificial agents. To enhance off-policy performance, we propose a simulation technique involving interactions across the entire agent space and simulated decision-makers. Our learning strategy yields significant OPE gains, e.g., improving prediction accuracy in the top 15\% challenging cases by 7.1%.

MURI: High-Quality Instruction Tuning Datasets for Low-Resource Languages via Reverse Instructions

Abdullatif Köksal, Marion Thaler, Ayyoob Imani, Ahmet Üstün, Anna Korhonen, Hinrich Schütze

Instruction tuning enhances large language models (LLMs) by aligning them with human preferences across diverse tasks. Traditional approaches to create instruction tuning datasets face serious challenges for low-resource languages due to their dependence on data annotation. This work introduces a novel method, Multilingual Reverse Instructions (MURI), which generates high-quality instruction tuning datasets for low-resource languages without requiring human annotators or pre-existing multilingual models. Utilizing reverse instructions and a translation pipeline, MURI produces instruction-output pairs from existing human-written texts in low-resource languages. This method ensures cultural relevance and diversity by sourcing texts from different native domains and applying filters to eliminate inappropriate content. Our dataset, MURI-IT, includes more than 2 million instruction-output pairs across 200 languages. Evaluation by native speakers and fine-tuning experiments with mT5 models demonstrate the approach's effectiveness for both NLU and open-ended generation. We publicly release datasets and models at https://github.com/akoksal/muri.

(Perhaps) Beyond Human Translation: Harnessing Multi-Agent Collaboration for Translating Ultra-Long Literary Texts

Minghao Wu, Jiahao Xu, Yulin Yuan, Gholamreza Haffari, Longyue Wang, Weihua Luo, Kaifu Zhang

Literary translation remains one of the most challenging frontiers in machine translation due to the complexity of capturing figurative language, cultural nuances, and unique stylistic elements. In this work, we introduce TransAgents, a novel multi-agent framework that simulates the roles and collaborative practices of a human translation company, including a CEO, Senior Editor, Junior Editor, Translator, Localization Specialist, and Proofreader. The translation process is divided into two stages: a preparation stage where the team is assembled and comprehensive translation guidelines are drafted, and an execution stage that involves sequential translation, localization, proofreading, and a final quality check. Furthermore, we propose two innovative evaluation strategies: Monolingual Human Preference (MHP), which evaluates translations based solely on target language quality and cultural appropriateness, and Bilingual LLM Preference (BLP), which leverages large language models like gpt-4 for direct text comparison. Although TransAgents achieves lower d-BLEU scores, due to the limited diversity of references, its translations are significantly better than those of other baselines and are preferred by both human evaluators and LLMs over traditional human references and gpt-4 translations. Our findings highlight the potential of multi-agent collaboration in enhancing translation quality, particularly for longer texts.

Continual Pre-training on Character-Level Noisy Texts Makes Decoder-based Language Models Robust Few-shot Learners

Takeshi Kojima, Yutaka Matsuo, Yusuke Iwasawa

Recent decoder-based pre-trained language models (PLMs) generally use subword tokenizers. However, adding character-level perturbations drastically changes the delimitation of texts by the tokenizers, leading to the vulnerability of PLMs. This study proposes a method of continual pre-training to convert decoder-based PLMs with subword tokenizers into perturbation-robust few-shot in-context learners. Our method continually trains decoder-based PLMs to predict the next tokens conditioning on artificially created character-level noisy texts. Since decoder-based language models are auto-regressive, we skip noised words from the target optimization. In addition, to maintain the same word prediction performance under noisy text as clean text, our method employs word distribution matching between the original PLMs and training models. We conducted experiments on various subword-based PLMs, including GPT2, Pythia, Mistral, Gemma2, and Llama3, ranging from 1B to 8B parameters. The results show that our method consistently improves downstream task performance for texts containing artificial noise and actual typos or spelling errors with few-shot in-context learning settings.

DARE: Diverse Visual Question Answering with Robustness Evaluation

Hannah Sterz, Jonas Pfeiffer, Ivan Vulic

Vision Language Models (VLMs) extend remarkable capabilities of text-only large language models and vision-only models, being able to learn from and process multi-modal vision-text input. While modern VLMs perform well on a number of standard image classification and image-text matching tasks, they still struggle with a number of crucial vision-language (VL) reasoning abilities such as counting and spatial reasoning. Moreover, while they might be very brittle to small variations in instructions and/or evaluation protocols, existing benchmarks fail to evaluate their robustness (or rather the lack of it). In order to couple challenging VL scenarios with comprehensive robustness evaluation, we introduce DARE, Diverse Visual Question Answering with Robustness Evaluation, a carefully created and curated multiple-choice VQA benchmark. DARE evaluates VLM performance on five diverse categories and includes four robustness-oriented evaluations based on the variations of: prompts, the subsets of answer options, the output format and the number of correct answers. Among a spectrum of other findings, we report that state-of-the-art VLMs still struggle with questions in most categories and are unable to consistently deliver their peak performance across the tested robustness evaluations. Consequently, our work calls for the systematic addition of robustness evaluations in future VLM research.

Explanatory Summarization with Discourse-Driven Planning

Dongqi Liu, Xi Yu, Vera Demberg, Mirella Lapata

Lay summaries for scientific documents typically include explanations to help readers grasp sophisticated concepts or arguments. However, current automatic summarization methods do not explicitly model explanations, which makes it difficult to align the proportion of explanatory content with human-written summaries. In this paper, we present a plan-based approach that leverages discourse frameworks to organize summary generation and guide explanatory sentences by prompting responses to the plan. Specifically, we propose two discourse-driven planning strategies, where the plan is conditioned as part of the input or part of the output prefix, respectively. Empirical experiments on three lay summarization datasets show that our approach outperforms existing state-of-the-art methods in terms of summary quality, and it enhances model robustness, controllability and mitigates hallucination.

Contextualized Evaluations: Taking the Guesswork Out of Language Model Evaluations

Chaitanya Malaviya, Joseph Chee Chang, Dan Roth, Mohit Iyyer, Mark Yatskar, Kyle Loa

Language model users often issue queries that lack specification, where the context under which a query was issued, such as the user's identity, the query's intent, and the criteria for a response to be useful is not explicit. For instance, a good response to a subjective query like “What book should I read next?” would depend on the user’s preferences, and a good response to an open-ended query like “How do antibiotics work against bacteria?” would depend on the user's expertise. This makes evaluation of responses to such queries an ill-posed task, as evaluators must make uninformed judgements about the quality of a response. To remedy this, we present contextualized evaluations, a protocol that synthetically constructs context surrounding an underspecified query and provides it during evaluation. We find that the presence of context can 1) change conclusions drawn from evaluation, even flipping win rates between model pairs, 2) nudge evaluators to make fewer judgments based on surface-level criteria of responses, like style, and 3) provide new insights about model behavior in different contexts. Specifically, we discover an implicit bias towards WEIRD contexts in default responses and find that models are not equally sensitive to following different contexts.

Virtual Presentation

Session 12: V-Presentations (Wednesday 30 July, 11:00-12:30,)

AI4Reading: Chinese Audiobook Interpretation System Based on Multi-Agent Collaboration

Minjiang Huang, Jipeng Qiang, Yi Zhu, Chaowei Zhang, Xiangyu Zhao, Kui Yu

Audiobook interpretations are attracting increasing attention, as they provide accessible and in-depth analyses of books that offer readers practical insights and intellectual inspiration. However, their manual creation process remains time-consuming and resource-intensive. To address this challenge, we propose AI4Reading, a multi-agent collaboration system leveraging large language models (LLMs) and speech synthesis technology to generate podcast-like audiobook interpretations. The system is designed to meet three key objectives: accurate content preservation, enhanced comprehensibility, and a logical narrative structure. To achieve these goals, We develop a framework composed of 11 specialized agents—including topic analysts, case analysts, editors, a narrator, and proofreaders—that work in concert to explore themes, extract real-world cases, refine content organization, and synthesize natural spoken language. By comparing expert interpretations with our system's output, the results show that although AI4Reading still has a gap in speech generation quality, the generated interpretative scripts are simpler and more accurate. The code of AI4Reading is publicly accessible \footnote{https://github.com/9624219/AI4reading}, with a demonstration video available \footnote{https://youtu.be/XCLAsRI9v2k}.

DocAgent: A Multi-Agent System for Automated Code Documentation Generation

Dayu Yang, Antoine Simoulin, Xin Qian, Xiaoyi Liu, Yuwei Cao, Zhaopu Teng, Grey Yang

High-quality code documentation is crucial for software development especially in the era of AI. However, generating it automatically using Large Language Models (LLMs) remains challenging, as existing approaches often produce incomplete, unhelpful, or factually incorrect outputs. We introduce DocAgent, a novel multi-agent collaborative system using topological code processing for incremental context building. Specialized agents (Reader, Searcher, Writer, Verifier, Orchestrator) then collaboratively generate documentation. We also propose a multi-faceted evaluation framework assessing Completeness, Helpfulness, and Truthfulness. Comprehensive experiments show DocAgent significantly outperforms baselines consistently. Our ablation study confirms the vital role of the topological processing order. DocAgent offers a robust approach for reliable code documentation generation in complex and proprietary repositories.

AI2Agent: An End-to-End Framework for Deploying AI Projects as Autonomous Agents

Jiaxiang Chen, Jingwei Shi, Lei Gan, Jiale Zhang, Qingyu Zhang, , Pang Xin, Zhucong Li, Xu Yinghui

As AI technology advances, it is driving innovation across industries, increasing the demand for scalable AI project deployment. However, deployment remains a critical challenge due to complex environment configurations, dependency conflicts, cross-platform adaptation, and debugging difficulties, which hinder automation and adoption. This paper introduces AI2Agent, an end-to-end framework that automates AI project deployment through guideline-driven execution, self-adaptive debugging, and case \\& solution accumulation. AI2Agent dynamically analyzes deployment challenges, learns from past cases, and iteratively refines its approach, significantly reducing human intervention. To evaluate its effectiveness, we conducted experiments on 30 AI deployment cases, covering TTS, text-to-image generation, image editing, and other AI applications. Results show that AI2Agent significantly reduces deployment time and improves success rates. The code\footnote{} and demo video\footnote{} are now publicly accessible.

Tree-of-Evolution: Tree-Structured Instruction Evolution for Code Generation in Large Language Models

Ziyang Luo, Kaixin Li, Hongzhan Lin, Yuchen Tian, Mohan Kankanhalli, Jing Ma

Data synthesis has become a crucial research area in large language models (LLMs), especially for generating high-quality instruction fine-tuning data to enhance downstream performance. In code generation, a key application of LLMs, manual annotation of code instruction data is costly. Recent methods, such as Code Evol-Instruct and OSS-Instruct, leverage LLMs to synthesize large-scale code instruction data, significantly improving LLM coding capabilities. However, these approaches face limitations due to unidirectional synthesis and randomness-driven generation, which restrict data quality and diversity. To overcome these challenges, we introduce Tree-of-Evolution (ToE), a novel framework that models code instruction synthesis process with a tree structure, exploring multiple evolutionary paths to alleviate the constraints of unidirectional generation. Additionally, we propose optimization-driven evolution, which refines each generation step based on the quality of the previous iteration. Experimental results across five widely-used coding benchmarks—HumanEval, MBPP, EvalPlus, LiveCodeBench, and BigCodeBench—demonstrate that base models fine-tuned on just 75k data synthesized by our method achieve comparable or superior performance to the state-of-the-art open-weight Code LLM, Qwen2.5-Coder-Instruct, which was fine-tuned on millions of samples.

Improve Vision Language Model Chain-of-thought Reasoning

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, Yiming Yang

Chain-of-thought (CoT) reasoning in vision language models (VLMs) is crucial for improving interpretability and trustworthiness. However, current training recipes often relying on datasets dominated by short annotations with minimal rationales. In this work, we show that training VLM on short answers leads to poor generalization on reasoning tasks that require more detailed explanations. To address this limitation, we propose a two-stage post-training strategy that extends the usage of short answer data for enhanced CoT reasoning. First, we augment short answers with CoT reasoning generated by GPT-4o, enhancing the VLM's CoT capabilities through fine-tuning. Second, we leverage short answers as outcome rewards for reinforcement learning. Specifically, short answers are used as correctness indicators to construct positive (correct) and negative (incorrect) pairs from model-generated reasoning chains. These pairs are then used to calibrate the model's reasoning via Direct Preference Optimization. Our experiments show significant improvements in CoT reasoning on benchmark datasets, along with enhanced generalization to direct answer prediction. This work provides a critical data resource for VLM CoT training and demonstrates the effectiveness of outcome rewards for multimodal models post-training.

Battling against Tough Resister: Strategy Planning with Adversarial Game for Non-collaborative Dialogues

Haiyang Wang, Zhiliang Tian, Yuchen Pan, Xin Song, Xin Niu, Minlie Huang, Bin Zhou

Non-collaborative dialogue involves two participants with conflicting interests engaging in a multi-round dialogue to achieve their own goals. Strategy planning is the key to guiding both participants towards a consensus. Most LLMs-based methods use stimulus prompts or external strategy planners for strategy planning. However, stimulus prompts fail to teach LLMs to plan dialogue strategies explicitly. Moreover, training external strategy planners doesn't fully account for adversarial interactions, thereby limiting their effectiveness against tough resisters. In this paper, to mitigate the above issues, we propose \textbf{GAIA}, a \textbf{G}ame-based \textbf{A}dversarial self-play \textbf{I}nter\textbf{A}ctive training paradigm, which constructs an adversarial two-player (a persuader and a resister) zero-sum game and guides the game to approximate Nash Equilibrium (NE) via reinforcement learning (RL) for the non-collaborative dialogues. First, we design a Chain-of-Mind prompt to reason the resister's dialogue act step-by-step to plan the persuasive strategies. Secondly, to adversarially improve the persuader, we construct diverse resistant planners and theoretically improve the persuader's optimal lower bound. Finally, we iteratively optimise their policies via adversarial self-play interactive RL and design an \epsilon-NE verification algorithm to approximate the game's NE. Experiments on three datasets show that our model obtains state-of-the-art performance.

VReST: Enhancing Reasoning in Large Vision-Language Models through Tree Search and Self-Reward Mechanism

Congzhi Zhang, Jiawei Peng, Zhenglin Wang, Yilong Lai, Haowen Sun, Heng Chang, Fei Ma, Weijiang Yu

Large Vision-Language Models (LVLMs) have shown exceptional performance in multimodal tasks, but their effectiveness in complex visual reasoning is still constrained, especially when employing Chain-of-Thought prompting techniques. In this paper, we propose VReST, a novel training-free approach that enhances Reasoning in LVLMs through Monte Carlo Tree Search and Self-Reward mechanisms. VReST meticulously traverses the reasoning landscape by establishing a search tree, where each node encapsulates a reasoning step, and each path delineates a comprehensive reasoning sequence. Our innovative multimodal Self-Reward mechanism assesses the quality of reasoning steps by integrating the utility of sub-questions, answer correctness, and the relevance of vision-language clues, all without the need for additional models. VReST surpasses current prompting methods and secures state-of-the-art performance across three multimodal mathematical reasoning benchmarks. Furthermore, it substantiates the efficacy of test-time scaling laws in multimodal tasks, offering a promising direction for future research.

BELLE: A Bi-Level Multi-Agent Reasoning Framework for Multi-Hop Question Answering

Taolin Zhang, Dongyang Li, Qizhou Chen, Chengyu Wang, Xiaofeng He

Multi-hop question answering (QA) involves finding multiple relevant passages and performing step-by-step reasoning to answer complex questions. Previous works on multi-hop QA employ specific methods from different modeling perspectives based on large language models (LLMs), regardless of the question types. In this paper, we first conduct an in-depth analysis of public multi-hop QA benchmarks, dividing the questions into four types and evaluating five types of cutting-edge methods for multi-hop QA: Chain-of-Thought (CoT), Single-step, Iterative-step, Sub-step, and Adaptive-step. We find that different types of multi-hop questions have varying degrees of sensitivity to different types of methods. Thus, we propose a Bi-levEL muLti-agEnt reasoning (BELLE) framework to address multi-hop QA by specifically focusing on the correspondence between question types and methods, where each type of method is regarded as an "operator" by prompting LLMs differently. The first level of BELLE includes multiple agents that debate to obtain an executive plan of combined "operators" to address the multi-hop QA task comprehensively. During the debate, in addition to the basic roles of affirmative debater, negative debater, and judge, at the second level, we further leverage fast and slow debaters to monitor whether changes in viewpoints are reasonable. Extensive experiments demonstrate that BELLE significantly outperforms strong baselines in various datasets. Additionally, the model consumption of BELLE is higher cost-effectiveness than that of single models in more complex multi-hop QA scenarios.

A Triple-View Framework for Fine-Grained Emotion Classification with Clustering-Guided Contrastive Learning

Junqing Gong, Binhan Yang, Wei Shen

Fine-grained emotion classification (FEC) aims to analyze speakers' utterances and distinguish dozens of emotions with subtle differences, allowing for a more nuanced understanding of human emotional states. However, compared to traditional coarse-grained emotion classification, two difficulties arise as the granularity of emotions becomes finer, i.e., the presence of closely confusable emotions which are hard to distinguish, and the biased performance caused by long-tailed emotions. Although addressing both difficulties is vital to FEC, previous studies have predominantly focused on dealing with only one of them. In this paper, we propose TACO, a novel triple-view framework that treats FEC as an instance-label (i.e., utterance-emotion) joint embedding learning problem to tackle both difficulties concurrently by considering three complementary views. Specifically, we design a clustering-guided contrastive loss, which incorporates clustering techniques to guide the contrastive learning process and facilitate more discriminative instance embeddings. Additionally, we introduce the emotion label description as a helpful resource to refine label embeddings and mitigate the poor performance towards under-represented (i.e., long-tailed) emotions. Extensive experiments on two widely-used benchmark datasets demonstrate that our proposed TACO achieves substantial and consistent improvements compared to other competitive baseline methods.

Beyond Facts: Evaluating Intent Hallucination in Large Language Models

Yijie Hao, Haofei Yu, Jiaxuan You

When exposed to complex queries containing multiple conditions, today’s large language models (LLMs) tend to produce responses that only partially satisfy the query while neglecting certain conditions. We, therefore, introduce the concept of Intent Hallucination, a phenomenon where LLMs either omit (failing to address certain parts) or misinterpret (responding to invented query parts) elements of the given query, leading to responses misaligned with the original query. To systematically evaluate intent hallucination, we introduce FAITHQA, a novel benchmark for intent hallucination that contains 20,068 problems, covering both query-only and retrieval-augmented generation (RAG) setups with varying topics and difficulty. FAITHQA is the first hallucination benchmark that goes beyond factual verification, tailored to identify the fundamental cause of intent hallucination. By evaluating various LLMs on FAITHQA, we find that (1) intent hallucination is a common issue even for state-of-the-art models, and (2) such a phenomenon stems from omission or misinterpretation of LLMs. To facilitate future research, we introduce an automatic LLM generation evaluation metric, named INTENT CONSTRAINT, for detecting intent hallucination. Human evaluation results demonstrate that INTENT CONSTRAINT is closer to human performance for intent hallucination compared to baselines.

CER: Confidence Enhanced Reasoning in LLMs

Ali Razghandi, Seyed Mohammad Hadi Hosseini, Mahdieh Soleymani Baghshah

Ensuring the reliability of Large Language Models (LLMs) in complex reasoning tasks remains a formidable challenge, particularly in scenarios that demand precise mathematical calculations and knowledge-intensive open-domain generation. In this work, we introduce an uncertainty-aware framework designed to enhance the accuracy of LLM responses by systematically incorporating model confidence at critical decision points. We propose an approach that encourages multi-step reasoning in LLMs and quantify the confidence of intermediate answers such as numerical results in mathematical reasoning and proper nouns in open-domain generation. Then, the overall confidence of each reasoning chain is evaluated based on confidence of these critical intermediate steps. Finally, we aggregate the answer of generated response paths in a way that reflects the reliability of each generated content (as opposed to self-consistency in which each generated chain contributes equally to majority voting). We conducted extensive experiments in five datasets, three mathematical datasets and two open-domain datasets, using four LLMs. The results consistently validate the effectiveness of our novel confidence-aggregation method, leading to an accuracy improvement of up to 7.4\% and 5.8\% over baseline approaches in math and open-domain generation tasks, respectively. Code is publicly available at https://github.com/sharif-ml-lab/CER.

Semantic-Eval : A Semantic Comprehension Evaluation Framework for Large Language Models Generation without Training

Shusheng Li, Jiale Li, Yifei Qu, Xinwei Shi, Yanliang Guo, Ziyi He, Yubo Wang, Wenjun Tan

With the increasing prominence of large language models (LLMs), evaluating their text-generation capabilities has become an essential research challenge. Although LLM-based evaluation methods exhibit robust performance, the inherent stochastic nature of the LLM generation process introduces a degree of uncertainty in alignment with human preferences. To address this limitation, we propose Semantic-Eval, the first training-free framework designed to assess LLM-generated text based on semantic understanding. This framework computes semantic similarity between pairwise texts to evaluate the interdependence of semantic units, integrating a graph-based weighting mechanism to account for the differential contributions of individual sentences. A pre-trained natural language inference (NLI) model is also incorporated to mitigate potential semantic relationship biases. We evaluate Semantic-Eval across eight datasets that encompass four common NLP tasks. The experimental results indicate that Semantic-Eval surpasses traditional N-gram and BERT-based evaluation metrics, aligning more closely with human judgments and demonstrating a higher correlation than smaller LLMs. However, it slightly lags behind GPT-4. Finally, we demonstrate the effectiveness of Semantic-Eval in evaluating the generation quality of 13 large language models. The code is publicly available at https://github.com/LssTry/Semantic-Eval.

SpeechFake: A Large-Scale Multilingual Speech Deepfake Dataset Incorporating Cutting-Edge Generation Methods

Wen Huang, Yanmei Gu, Zhiming Wang, Huijia Zhu, Yanmin Qian

As speech generation technology advances, the risk of misuse through deepfake audio has become a pressing concern, which underscores the critical need for robust detection systems. However, many existing speech deepfake datasets are limited in scale and diversity, making it challenging to train models that can generalize well to unseen deepfakes. To address these gaps, we introduce SpeechFake, a large-scale dataset designed specifically for speech deepfake detection. SpeechFake includes over 3 million deepfake samples, totaling more than 3,000 hours of audio, generated using 40 different speech synthesis tools. The dataset encompasses a wide range of generation techniques, including text-to-speech, voice conversion, and neural vocoder, incorporating the latest cutting-edge methods. It also provides multilingual support, spanning 46 languages. In this paper, we offer a detailed overview of the dataset's creation, composition, and statistics. We also present baseline results by training detection models on SpeechFake, demonstrating strong performance on both its own test sets and various unseen test sets. Additionally, we conduct experiments to rigorously explore how generation methods, language diversity, and speaker variation affect detection performance. We believe SpeechFake will be a valuable resource for advancing speech deepfake detection and developing more robust models for evolving generation techniques.

Enhancing Neural Machine Translation Through Target Language Data: A kNN-LM Approach for Domain Adaptation

Abudurexiti Reheman, Hongyu Liu, Junhao Ruan, Abudukeyumu Abudula, yingfeng luo, Tong Xiao, JingBo Zhu

Neural machine translation (NMT) has advanced significantly, yet challenges remain in adapting to new domains . In scenarios where bilingual data is limited, this issue is further exacerbated. To address this, we propose kNN-LM-NMT, a method that leverages semantically similar target language sentences in the kNN framework. Our approach generates a probability distribution over these sentences during decoding, and this distribution is then interpolated with the NMT model's distribution. Additionally, we introduce an n-gram-based approach to focus on similar fragments, enabling the model to avoid the noise introduced by the non-similar parts. To enhance accuracy, we further incorporate cross-lingual retrieval similarity to refine the kNN probability distribution. Extensive experiments on multi-domain datasets demonstrate significant performance improvements in both high-resource and low-resource scenarios. Our approach effectively extracts translation knowledge from limited target domain data, and well benefits from large-scale monolingual data for robust context representation.

Uni-Retrieval: A Multi-Style Retrieval Framework for STEM's Education

Yanhao Jia, Xinyi Wu, Li Hao, QinglinZhang, Yuxiao Hu, Shuai Zhao, Wenqi Fan

In AI-facilitated teaching, leveraging various query styles to interpret abstract text descriptions is crucial for ensuring high-quality teaching. However, current retrieval models primarily focus on natural text-image retrieval, making them insufficiently tailored to educational scenarios due to the ambiguities in the retrieval process. In this paper, we propose a diverse expression retrieval task tailored to educational scenarios, supporting retrieval based on multiple query styles and expressions. We introduce the STEM Education Retrieval Dataset (SER), which contains over 24,000 query pairs of different styles, and the Uni-Retrieval, an efficient and style-diversified retrieval vision-language model based on prompt tuning. Uni-Retrieval extracts query style features as prototypes and builds a continuously updated Prompt Bank containing prompt tokens for diverse queries. This bank can updated during test time to represent domain-specific knowledge for different subject retrieval scenarios. Our framework demonstrates scalability and robustness by dynamically retrieving prompt tokens based on prototype similarity, effectively facilitating learning for unknown queries. Experimental results indicate that Uni-Retrieval outperforms existing retrieval models in most retrieval tasks.

Enhancing Safe and Controllable Protein Generation via Knowledge Preference Optimization

Yuhao Wang, Keyan Ding, Kehua Feng, Zeyuan Wang, Ming Qin, Xiaotong Li, Qiang Zhang, Huajun Chen

Protein language models have emerged as powerful tools for sequence generation, offering substantial advantages in functional optimization and *denovo* design. However, these models also present significant risks of generating harmful protein sequences, such as those that enhance viral transmissibility or evade immune responses. These concerns underscore critical biosafety and ethical challenges. To address these issues, we propose a Knowledge-guided Preference Optimization (KPO) framework that integrates prior knowledge via a Protein Safety Knowledge Graph. This framework utilizes an efficient graph pruning strategy to identify preferred sequences and employs reinforcement learning to minimize the risk of generating harmful proteins. Experimental results demonstrate that KPO effectively reduces the likelihood of producing hazardous sequences while maintaining high functionality, offering a robust safety assurance framework for applying generative models in biotechnology.

Can Large Language Models Understand Internet Buzzwords Through User-Generated Content

Chen Huang, Junkai Luo, Xinzuo Wang, Wenqiang Lei, Jiancheng Lv

The massive user-generated content (UGC) available in Chinese social media is giving rise to the possibility of studying internet buzzwords. In this paper, we study if large language models (LLMs) can generate accurate definitions for these buzzwords based on UGC as examples. Our work serves a threefold contribution. First, we introduce CHEER, the first dataset of Chinese internet buzzwords, each annotated with a definition and relevant UGC. Second, we propose a novel method, called RESS, to effectively steer the comprehending process of LLMs to produce more accurate buzzword definitions, mirroring the skills of human language learning. Third, with CHEER, we benchmark the strengths and weaknesses of various off-the-shelf definition generation methods and our RESS. Our benchmark demonstrates the effectiveness of RESS while revealing a crucial shared challenge: comprehending unseen buzzwords and leveraging sufficient, high-quality UGC to facilitate this comprehension. In this paper, we believe our work lays the groundwork for future advancements in LLM-based definition generation. Our dataset and code will be openly released.

MasRouter: Learning to Route LLMs for Multi-Agent Systems

Yanwei Yue, Guibin Zhang, Boyang Liu, Guancheng Wan, Kun Wang, Dawei Cheng, Yiyan Qi

Multi-agent systems (MAS) powered by Large Language Models (LLMs) have been demonstrated to push the boundaries of LLM capabilities, yet they often incur significant costs and face challenges in dynamic LLM selection. Current LLM routing methods effectively reduce overhead in single-agent scenarios by customizing LLM selection for each query, but they overlook the critical decisions regarding collaboration modes and agent roles in MAS. In response to this challenge, we first introduce the problem of \textbf{Multi-Agent System Routing (MASR)}, which integrates all components of MAS into a unified routing framework. Toward this goal, we propose MasRouter, the first high-performing, cost-effective, and inductive \textbf{MASR} solution. MasRouter employs collaboration mode determination, role allocation, and LLM routing through a cascaded controller network, progressively constructing a MAS that balances effectiveness and efficiency. Extensive experiments demonstrate that MasRouter is \textbf{(1) high-performing}, achieving a 1.8 improvement over the state-of-the-art method on MBPP; \textbf{(2) economical}, reducing overhead by up to 52.07 compared to SOTA methods on HumanEval; and \textbf{(3) plug-and-play}, seamlessly integrating with mainstream MAS frameworks, reducing overhead by 17.21 via customized routing.

VLMInferSlow: Evaluating the Efficiency Robustness of Large Vision-Language Models as a Service

Xiasi Wang, Tianliang Yao, Simin Chen, Runqi Wang, Lei YE, Kuofeng Gao, Yi Huang, Yuan Yao

Vision-Language Models (VLMs) have demonstrated great potential in real-world applications. While existing research primarily focuses on improving their accuracy, the efficiency remains underexplored. Given the real-time demands of many applications and the high inference overhead of VLMs, efficiency robustness is a critical issue. However, previous studies evaluate efficiency robustness under unrealistic assumptions, requiring access to the model architecture and parameters---an impractical scenario in ML-as-a-service settings, where VLMs are deployed via inference APIs. To address this gap, we propose VLMInferSlow, a novel approach for evaluating VLM efficiency robustness in a realistic black-box setting. VLMInferSlow incorporates fine-grained efficiency modeling tailored to VLM inference and leverages zero-order optimization to search for adversarial examples. Experimental results show that VLMInferSlow generates adversarial images with imperceptible perturbations, increasing the computational cost by up to 128.47%. We hope this research raises the community's awareness about the efficiency robustness of VLMs.

The Role of Deductive and Inductive Reasoning in Large Language Models

Chengkun Cai, Xu Zhao, Haoliang Liu, Zhongyu Jiang, Tianfang Zhang, Zongkai Wu, Jenq-Neng Hwang, Serge Belongie, Lei Li

Large Language Models (LLMs) have demonstrated impressive capabilities in reasoning tasks, yet their reliance on static prompt structures and limited adaptability to complex scenarios remains a major challenge. In this paper, we propose the **Deductive and Inductive (DID)** method, a novel framework that enhances LLM reasoning by dynamically integrating both deductive and inductive reasoning approaches. Drawing from cognitive science principles, DID implements a dual-metric complexity evaluation system that combines Littlestone dimension and information entropy to precisely assess task difficulty and guide decomposition strategies. DID enables the model to progressively adapt its reasoning pathways based on problem complexity, mirroring human cognitive processes. We evaluate DID's effectiveness across multiple benchmarks, including the AIW, MR-GSM8K, and our custom Holiday Puzzle dataset for temporal reasoning. Our results demonstrate great improvements in reasoning quality and solution accuracy - achieving 70.3% accuracy on AIW (compared to 62.2% for Tree of Thought), while maintaining lower computational costs.

Table-Critic: A Multi-Agent Framework for Collaborative Criticism and Refinement in Table Reasoning

Peiying Yu, Guoxin Chen, Jingjing Wang

Despite the remarkable capabilities of large language models (LLMs) in various reasoning tasks, they still struggle with table reasoning tasks, particularly in maintaining consistency throughout multi-step reasoning processes. While existing approaches have explored various decomposition strategies, they often lack effective mechanisms to identify and correct errors in intermediate reasoning steps, leading to cascading error propagation. To address these issues, we propose Table-Critic, a novel multi-agent framework that facilitates collaborative criticism and iterative refinement of the reasoning process until convergence to correct solutions. Our framework consists of four specialized agents: a Judge for error identification, a Critic for comprehensive critiques, a Refiner for process improvement, and a Curator for pattern distillation. To effectively deal with diverse and unpredictable error types, we introduce a self-evolving template tree that systematically accumulates critique knowledge through experience-driven learning and guides future reflections. Extensive experiments have demonstrated that Table-Critic achieves substantial improvements over existing methods, achieving superior accuracy and error correction rates while maintaining computational efficiency and lower solution degradation rate.

Soundwave: Less is More for Speech-Text Alignment in LLMs

Yuhao Zhang, Zhiheng Liu, Fan Bu, Ruiyu Zhang, Benyou Wang, Haizhou Li

Existing end-to-end speech large language models (LLMs) usually rely on large-scale annotated data for training, while data-efficient training has not been discussed in depth. We focus on two fundamental problems between speech and text: the representation space gap and sequence length inconsistency. We propose Soundwave, which utilizes an efficient training strategy and a novel architecture to address these issues. Results show that Soundwave outperforms other advanced speech LLMs in speech translation and AIR-Bench speech tasks with only a fraction of the training data. Further analysis shows that Soundwave still retains its intelligence during conversation.

Proxy-Driven Robust Multimodal Sentiment Analysis with Incomplete Data

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, Ning An

Multimodal Sentiment Analysis (MSA) with incomplete data has gained significant attention recently. Existing studies focus on optimizing model structures to handle modality missingness, but models still face challenges in robustness when dealing with uncertain missingness. To this end, we propose a data-centric robust multimodal sentiment analysis method, Proxy-Driven Robust Multimodal Fusion (P-RMF). First, we map unimodal data to the latent space of Gaussian distributions to capture core features and structure, thereby learn stable modality representation. Then, we combine the quantified inherent modality uncertainty to learn stable multimodal joint representation (i.e., proxy modality), which is further enhanced through multi-layer dynamic cross-modal injection to increase its diversity. Extensive experimental results show that P-RMF outperforms existing models in noise resistance and achieves state-of-the-art performance on multiple benchmark datasets. Code will be available at https://github.com/***/P-RMF.

ChartLens: Fine-grained Visual Attribution in Charts

Manan Suri, Puneet Mathur, Nedim Lipka, Franck Dernoncourt, Ryan A. Rossi, Dinesh Manocha

The growing capabilities of multimodal large language models (MLLMs) have advanced tasks like chart understanding. However, these models often suffer from hallucinations, where generated text sequences conflict with the provided visual data. To address this, we introduce Post-Hoc Visual Attribution for Charts, which identifies fine-grained chart elements that validate a given chart-associated response. We propose ChartLens, a novel chart attribution algorithm that uses segmentation-based techniques to identify chart objects and employs set-of-marks prompting with MLLMs for fine-grained visual attribution. Additionally, we present ChartVA-Eval, a benchmark with synthetic and real-world charts from diverse domains like finance, policy, and economics, featuring fine-grained attribution annotations. Our evaluations show that ChartLens improves fine-grained attributions by 26-66%.

Beyond Logits: Aligning Feature Dynamics for Effective Knowledge Distillation

Guoqiang Gong, Jiaxing Wang, Jin Xu, Deping Xiang, Zicheng Zhang, Leqi Shen, Yifeng Zhang, JunhuaShu, ZhaolongXing, Zhen Chen, Pengzhang Liu, Ke Zhang

Knowledge distillation (KD) compresses large language models (LLMs), known as teacher models, into lightweight versions called student models, enabling efficient inference and downstream applications. However, prevailing approaches accomplish this by predominantly focusing on matching the final output distributions of student/teacher models. Drawing on the perspective that transformers can be viewed as discretizing ordinary differential equation (ODEs) on integer time steps (corresponding to layer indices), where intermediate features evolve across layers, we argue that effective KD requires aligning the entire feature dynamics between teacher and student models, which we call feature dynamics distillation (FDD). This alignment involves matching both the feature trajectory and its first-order derivative, rather than just the final states. Our approach extends the original KD objective with two additional loss terms: layer-wise feature KD, which matches discretized feature trajectory, and layer feature delta KD, which matches first-order changes in features across adjacent layers. Extensive experiments on various tasks validate the effectiveness of our distillation method.

Curiosity-Driven Reinforcement Learning from Human Feedback

Haoran Sun, Yekun Chai, Shuohuan Wang, Yu Sun, Hua Wu, Haifeng Wang

Reinforcement learning from human feedback (RLHF) has proven effective in aligning large language models (LLMs) with human preferences, but often at the cost of reduced output diversity. This trade-off between diversity and alignment quality remains a significant challenge. Drawing inspiration from curiosity-driven exploration in reinforcement learning, we introduce curiosity-driven RLHF (CD-RLHF), a framework that incorporates intrinsic rewards for novel states, alongside traditional sparse extrinsic rewards, to optimize both output diversity and alignment quality. We demonstrate the effectiveness of CD-RLHF through extensive experiments on a range of tasks, including text summarization and instruction following. Our approach achieves significant gains in diversity on multiple diversity-oriented metrics while maintaining alignment with human preferences comparable to standard RLHF. We will make our code publicly available.

A Dual-Mind Framework for Strategic and Expressive Negotiation Agent

Yutong Liu, Lida Shi, Rui Song, Hao Xu

Negotiation agents need to influence the attitudes or intentions of users to reach a consensus. Strategy planning and expressive optimization are crucial aspects of effective negotiations. However, previous studies have typically focused on only one of these aspects, neglecting the fact that their combined synergistic effect can lead to better performance. Inspired by the dual-process theory in human cognition, we propose a Dual-Mind Negotiation Agent (DMNA) framework. This framework integrates an intuitive module for rapid, experience-based response and a deliberative module for slow, expression optimization. The intuitive module is trained using Monte Carlo Tree Search (MCTS) and Direct Preference Optimization (DPO), enabling it to make suitable strategic planning and expression. The deliberative module employs a multifaceted reflexion mechanism to enhance the quality of expression. Experiments conducted on negotiation datasets confirm that DMNA achieves state-of-the-art results, demonstrating an enhancement in the negotiation ability of agents.

Revisiting Scaling Laws for Language Models: The Role of Data Quality and Training Strategies

Zhengyu Chen, Siqi Wang, Teng Xiao, Yudong Wang, Shiqi Chen, Xunliang Cai, Junxian He, Jingang Wang

Traditional scaling laws in natural language processing suggest that increasing model size and training data enhances performance. However, recent studies reveal deviations, particularly in large language models, where performance improvements decelerate—a phenomenon known as sub-scaling. This paper revisits these scaling laws by examining the impact of data quality and training strategies on model performance. Through extensive empirical analysis of over 400 models, we identify high data density and non-optimal resource allocation as key factors contributing to sub-scaling. High data density leads to diminishing returns due to redundant information, while optimal resource allocation is crucial for sustained performance improvements. We propose a sub-optimal scaling law that better predicts performance in sub-scaling regimes, highlighting the importance of data quality and diversity.

Document-Level Event-Argument Data Augmentation for Challenging Role Types

Joseph Gatto, Omar Sharif, Parker Seegmiller, Sarah Masud Preum

Event Argument Extraction (EAE) is a daunting information extraction problem --- with significant limitations in few-shot cross-domain (FSCD) settings. A common solution to FSCD modeling is data augmentation. Unfortunately, existing augmentation methods are not well-suited to a variety of real-world EAE contexts, including (i) modeling long documents (documents with over 10 sentences), and (ii) modeling challenging role types (i.e., event roles with little to no training data and semantically outlying roles). We introduce two novel LLM-powered data augmentation methods for generating extractive document-level EAE samples using zero in-domain training data. We validate the generalizability of our approach on four datasets --- showing significant performance increases in low-resource settings. Our highest performing models provide a 13-pt increase in F1 score on zero-shot role extraction in FSCD evaluation.

ReflectDiffu: Reflect between Emotion-intent Contagion and Mimicry for Empathetic Response Generation via a RL-Diffusion Framework

Jiahao Yuan, Zixiang Di, Zhiqing Cui, Guisong Yang, Usman Naseem

Empathetic response generation necessitates the integration of emotional and intentional dynamics to foster meaningful interactions. Existing research either neglects the intricate interplay between emotion and intent, leading to suboptimal controllability of empathy, or resorts to large language models (LLMs), which incur significant computational overhead. In this paper, we introduce ReflectDiffu, a lightweight and comprehensive framework for empathetic response generation. This framework incorporates emotion contagion to augment emotional expressiveness and employs an emotion-reasoning mask to pinpoint critical emotional elements. Additionally, it integrates intent mimicry within reinforcement learning for refinement during diffusion. By harnessing an intent twice reflect mechanism of Exploring-Sampling-Correcting, ReflectDiffu adeptly translates emotional decision-making into precise intent actions, thereby addressing empathetic response misalignments stemming from emotional misrecognition. Through reflection, the framework maps emotional states to intents, markedly enhancing both response empathy and flexibility. Comprehensive experiments reveal that ReflectDiffu outperforms existing models regarding relevance, controllability, and informativeness, achieving state-of-the-art results in both automatic and human evaluations.

Do LLMs Understand Dialogues? A Case Study on Dialogue Acts

Ayesha Qamar, Jonathan Tong, Ruihong Huang

Recent advancements in NLP, largely driven by Large Language Models (LLMs), have significantly improved performance on an array of tasks. However, Dialogue Act (DA) classification remains challenging, particularly in the fine-grained 50-class, multiparty setting. This paper investigates the root causes of LLMs' poor performance in DA classification through a linguistically motivated analysis. We identify three key pre-tasks essential for accurate DA prediction: Turn Management, Communicative Function Identification, and Dialogue Structure Prediction. Our experiments reveal that LLMs struggle with these fundamental tasks, often failing to outperform simple rule-based baselines. Additionally, we establish a strong empirical correlation between errors in these pre-tasks and DA classification failures. A human study further highlights the significant gap between LLM and human-level dialogue understanding. These findings indicate that LLMs' shortcomings in dialogue comprehension hinder their ability to accurately predict DAs, highlighting the need for improved dialogue-aware training approaches.

Generative Reward Modeling via Synthetic Criteria Preference Learning

xiaobo liang, Haoke Zhang, Juntao Li, Kehai Chen, Qiaoming Zhu, Min Zhang

Generative Reward Models (GenRMs) leverage synthesized Chains of Thought (CoT) to reduce the need for massive labeled data, but this approach introduces risks of overoptimization due to the inability to guarantee the correctness of the CoTs. Identifying and optimizing unexpected behaviors within these synthesized CoT remains a challenge, as it heavily depends on precise annotations of intermediate behavior, similar to process supervision. In this work, we introduce a criteria-based preference tree for reward modeling, where each path in the tree represents a reasoning trajectory based on synthesized criteria. Crucially, each reasoning trajectory can be independently optimized through RL algorithm. These fine-grained process reward signals are derived from the inference-time computations and predefined rules, eliminating the need for human supervision. In experiments, SyncPL showed significant improvements over baselines on multiple human preference benchmarks. We further demonstrate that synthesized data can be learned using a long CoT format, analogous to an o1-like model, further enhancing performance while keeping stability and efficiency during training.

Revisiting LLMs as Zero-Shot Time Series Forecasters: Small Noise Can Break Large Models

Junwoo Park, Hyuck Lee, Dohyun Lee, Daehoon Gwak, Jaegul Choo

Large Language Models (LLMs) have shown remarkable performance across diverse tasks without domain-specific training, fueling interest in their potential for time-series forecasting. While LLMs have shown potential in zero-shot forecasting through prompting alone, recent studies suggest that LLMs lack inherent effectiveness in forecasting. Given these conflicting findings, a rigorous validation is essential for drawing reliable conclusions. In this paper, we evaluate the effectiveness of LLMs as zero-shot forecasters compared to state-of-the-art domain-specific models. Our experiments show that LLM-based zero-shot forecasters often struggle to achieve high accuracy due to their sensitivity to noise, underperforming even simple domain-specific models. We have explored solutions to reduce LLMs' sensitivity to noise in the zero-shot setting, but improving their robustness remains a significant challenge. Our findings suggest that rather than emphasizing zero-shot forecasting, a more promising direction would be to focus on fine-tuning LLMs to better process numerical sequences. Our experimental code is available at https://github.com/junwoopark92/revisiting-LLMs-zeroshot-forecaster.

Online Iterative Self-Alignment for Radiology Report Generation

Ting Xiao, Lei Shi, Yang Zhang, HaoFeng Yang, Zhe Wang, Chenjia Bai

Radiology Report Generation (RRG) is an important research topic for relieving radiologists' heavy workload. Existing RRG models mainly rely on supervised fine-tuning (SFT) based on different model architectures using data pairs of radiological images and corresponding radiologist-annotated reports. Recent research has shifted focus to post-training improvements, aligning RRG model outputs with human preferences using reinforcement learning (RL). However, the limited data coverage of high-quality annotated data poses risks of overfitting and generalization. This paper proposes a novel Online Iterative Self-Alignment (OISA) method for RRG that consists of four stages: self-generation of diverse data, self-evaluation for multi-objective preference data, self-alignment for multi-objective optimization and self-iteration for further improvement. Our approach allows for generating varied reports tailored to specific clinical objectives, enhancing the overall performance of the RRG model iteratively. Unlike existing methods, our framework significantly increases data quality and optimizes performance through iterative multi-objective optimization. Experimental results demonstrate that our method surpasses previous approaches, achieving state-of-the-art performance across multiple evaluation metrics.

Parameter-Aware Contrastive Knowledge Editing: Tracing and Rectifying based on Critical Transmission Paths

Songlin Zhai, Yuan Meng, Yuxin Zhang, Guilin Qi

Large language models (LLMs) have encoded vast amounts of knowledge in their parameters, but the acquired knowledge can sometimes be incorrect or outdated over time, necessitating rectification after pre-training. Traditional localized methods in knowledge-based model editing (KME) typically assume that knowledge is stored in particular intermediate layers. However, recent research suggests that these methods do not identify the optimal locations for parameter editing, as knowledge gradually accumulates across all layers in LLMs during the forward pass rather than being stored in specific layers. This paper, for the first time, introduces the concept of critical transmission paths into KME for parameter updating. Specifically, these paths capture the key information flows that significantly influence the model predictions for the editing process. To facilitate this process, we also design a parameter-aware contrastive rectifying algorithm that considers less important paths as contrastive examples. Experiments on two prominent datasets and three widely used LLMs demonstrate the superiority of our method in editing performance.

Many Heads Are Better Than One: Improved Scientific Idea Generation by A LLM-Based Multi-Agent System

Haoyang Su, Renqi Chen, SHIXIANG TANG, Zhenfei Yin, Xinzhe Zheng, Jinzhe Li, Biqing Qi, Qi Wu, Hui Li, Wanli Ouyang, Philip Torr, Bowen Zhou, Nanqing Dong

The rapid advancement of scientific progress requires innovative tools that can accelerate knowledge discovery. Although recent AI methods, particularly large language models (LLMs), have shown promise in tasks such as hypothesis generation and experimental design, they fall short of replicating the collaborative nature of real-world scientific practices, where diverse experts work together in teams to tackle complex problems. To address the limitations, we propose an LLM-based multi-agent system, i.e., Virtual Scientists (VIRSCI), designed to mimic the teamwork inherent in scientific research. VIRSCI organizes a team of agents to collaboratively generate, evaluate, and refine research ideas. Through comprehensive experiments, we demonstrate that this multi-agent approach outperforms the state-of-the-art method in producing novel scientific ideas. We further investigate the collaboration mechanisms that contribute to its tendency to produce ideas with higher novelty, offering valuable insights to guide future research and illuminating pathways toward building a robust system for autonomous scientific discovery. The code is available at https://github.com/open-sciencelab/Virtual-Scientists.

LLM-based Rumor Detection via Influence Guided Sample Selection and Game-based Perspective Analysis

Zhiliang Tian, jingyuan huang, Zejiang He, Zhen Huang, Menglong Lu, Linbo Qiao, Songzhu Mei, Yijie Wang, Dongsheng Li

Rumor detection on social media has become an emerging topic. Traditional deep learning-based methods model rumors based on content, propagation structure, or user behavior, but these approaches are constrained by limited modeling capacity and insufficient training corpora. Recent studies have explored using LLMs for rumor detection through supervised fine-tuning (SFT), but face two issues: 1) unreliable samples sometimes mislead the model learning; 2) the model only learns the most salient input-output mapping and skips in-depth analyses of the rumored content for convenience. To address these issues, we propose an SFT-based LLM rumor detection model with Influence guided Sample selection and Game-based multi-perspective Analysis (ISGA). Specifically, we first introduce the Influence Score (IS) to assess the impact of samples on model predictions and select samples for SFT. We also approximate IS via Taylor expansion to reduce computational complexity. Next, we use LLMs to generate in-depth analyses of news content from multiple perspectives and model their collaborative process for prediction as a cooperative game. Then we utilize the Shapley value to quantify the contribution of each perspective for selecting informative perspective analyses. Experiments show that ISGA excels existing SOTA on three datasets.

Can LLMs Generate High-Quality Test Cases for Algorithm Problems? TestCase-Eval: A Systematic Evaluation of Fault Coverage and Exposure

Zheyuan Yang, Zexi Kuang, Yilun Zhao

We introduce TestCase-Eval, a new benchmark for systematic evaluation of LLMs in test-case generation. TestCase-Eval includes 500 algorithm problems and 100,000 human-crafted solutions from the Codeforces platform. It focuses on two pivotal tasks: (1) Fault Coverage, which measures how well LLM-generated test sets probe diverse input scenarios and cover a wide range of potential failure modes. (2) Fault Exposure, which evaluates whether LLMs can craft a tailored test input that reveals a specific incorrect code implementation. We provide a comprehensive assessment of 19 state-of-the-art open-source and proprietary LLMs on TestCase-Eval, offering insights into their strengths and limitations in generating effective test cases for algorithm problems.

Should I Believe in What Medical AI Says? A Chinese Benchmark for Medication Based on Knowledge and Reasoning

Yue Wu, Yangmin Huang, Qianyun Du, Lixian Lai, Zhiyang He, Jiaxue Hu, Xiaodong Tao

Large language models (LLMs) show potential in healthcare but often generate hallucinations, especially when handling unfamiliar information. In medication, a systematic benchmark to evaluate model capabilities is lacking, which is critical given the high-risk nature of medical information. This paper introduces a Chinese benchmark aimed at assessing models in medication tasks, focusing on knowledge and reasoning across six datasets: indication, dosage and administration, contraindicated population, mechanisms of action, drug recommendation, and drug interaction. We evaluate eight closed-source and five open-source models to identify knowledge boundaries, providing the first systematic analysis of limitations and risks in proprietary medical models.

TestNUC: Enhancing Test-Time Computing Approaches through Neighboring Unlabeled Data Consistency

Henry Peng Zou, Zhengyao Gu, Yue Zhou, Yankai Chen, Weizhi Zhang, Liancheng Fang, Yibo Wang, Yangning Li, Kay Liu, Philip S. Yu

Test-time computing approaches, which leverage additional computational resources during inference, have been proven effective in enhancing large language model performance. This work introduces a novel, linearly scaling approach, TestNUC, that improves test-time predictions by leveraging the local consistency of neighboring unlabeled data-it classifies an input instance by considering not only the model's prediction on that instance but also on neighboring unlabeled instances. We evaluate TestNUC across eight diverse datasets, spanning intent classification, topic mining, domain discovery, and emotion detection, demonstrating its consistent superiority over baseline methods such as standard prompting and self-consistency. Furthermore, TestNUC can be seamlessly integrated with existing test-time computing approaches, substantially boosting their performance. Our analysis reveals that TestNUC scales effectively with increasing amounts of unlabeled data and performs robustly across different embedding models, making it practical for real-world applications. Our code is available at https://github.com/HenryPengZou/TestNUC.

Hidden in Plain Sight: Evaluation of the Deception Detection Capabilities of LLMs in Multimodal Settings

Md Messal Monem Miah, Adrita Anika, Xi Shi, Ruihong Huang

Detecting deception in an increasingly digital world is both a critical and challenging task. In this study, we present a comprehensive evaluation of the automated deception detection capabilities of Large Language Models (LLMs) and Large Multimodal Models (LMMs) across diverse domains. We assess the performance of both open-source and proprietary LLMs on three distinct datasets—real-life trial interviews (RLTD), instructed deception in interpersonal scenarios (MU3D), and deceptive reviews (OpSpam). We systematically analyze the effectiveness of different experimental setups for deception detection, including zero-shot and few-shot approaches with random or similarity-based in-context example selection. Our findings indicate that fine-tuned LLMs achieve state-of-the-art performance on textual deception detection, whereas LMMs struggle to fully leverage multimodal cues, particularly in real-world settings. Additionally, we analyze the impact of auxiliary features, such as non-verbal gestures, video summaries, and evaluate the effectiveness of different prompting strategies, such as direct label generation and post-hoc reasoning generation. Experiments unfold that reasoning-based predictions do not consistently improve performance over direct classification, contrary to the expectations.

CMHKF: Cross-Modality Heterogeneous Knowledge Fusion for Weakly Supervised Video Anomaly Detection

Shengping Song, Yongsen Zheng, Wuchun He, Guohua Wang

Weakly supervised video anomaly detection (WSVAD) presents a challenging task focused on detecting frame-level anomalies using only video-level labels. However, existing methods focus mainly on visual modalities, neglecting rich multi-modality information. This paper proposes a novel framework, Cross-Modality Heterogeneous Knowledge Fusion (CMHKF), that integrates cross-modality knowledge from video, audio, and text to improve anomaly detection and localization. To achieve adaptive cross-modality heterogeneous knowledge learning, we designed two components: Cross-Modality Video-Text Knowledge Alignment (CVKA) and Audio Modality Feature Adaptive Extraction (AFAE). They extract and aggregate features by exploring inter-modality correlations. By leveraging abundant cross-modality knowledge, our approach improves the discrimination between normal and anomalous segments. Extensive experiments on XD-Violence show our method significantly enhances accuracy and robustness in both coarse-grained and fine-grained anomaly detection.

12 Workshops

Overview

During the days of the workshops, **Registration** will be held from 08:00- 16:30.
Workshop Posters will be on Level 1 Hall L1 - L3.

Thursday, July 31, 2025

Level 1 1.34	Workshop 1: 6th AfricaNLP Workshop: Multilingual and Multicultural-aware LLMs	p.817
Level 2 Hall B	Workshop 2: The 12th Workshop on Argument Mining	p.818
Level 2 2.17	Workshop 3: 2nd Workshop on Natural Language Processing meets Climate Change	p.819
Level 2 2.31	Workshop 4: Eighth Workshop on Fact Extraction and VERification (FEVER)	p.820
Level 2 Hall C	Workshop 5: GEM2 Workshop: Generation, Evaluation & Metrics	p.821
Level 1 1.15-16	Workshop 6: The 19th Linguistic Annotation Workshop (LAW XIX)	p.822
Level 1 Hall N1	Workshop 7: The 4th Workshop on NLP for Positive Impact	p.823
Level 1 1.61-62	Workshop 8: REALM: First Workshop for Research on Agent Language Models	p.824
Level 1 1.14	Workshop 9: SDP 2025: The 5th Workshop on Scholarly Document Processing	p.825
Level 1 1.31-32	Workshop 10: The Third Workshop on Social Influence in Conversations (SICon 2025)	p.826
Level 2 2.44	Workshop 11: Slavic NLP-2025: The 10th Biennial Workshop on Slavic NLP	p.827
Level 2 2.15	Workshop 12: The 4th Table Representation Learning Workshop at ACL 2025	p.828
Gather	Workshop 13: The Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)	p.829
Level 1 1.85-86	Workshop 14: The 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)	p.830
Level 1 Hall M1	Workshop 15: 29th Conference on Computational Natural Language Learning (CoNLL)	p.831
Level 1 Hall N2	Workshop 16: The 22nd International Conference on Spoken Language Translation (IWSLT 2025)	p.832
Level 1 Hall M2	Workshop 17: The 19th Workshop on Semantic Evaluation (Semeval-2025)	p.833

Friday, August 1, 2025

Gather	Workshop 13: The Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)	p.829
Level 1 1.85-86	Workshop 14: The 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)	p.830
Level 1 Hall M1	Workshop 15: 29th Conference on Computational Natural Language Learning (CoNLL)	p.831
Level 1 Hall N2	Workshop 16: The 22nd International Conference on Spoken Language Translation (IWSLT 2025)	p.832
Level 1 Hall M2	Workshop 17: The 19th Workshop on Semantic Evaluation (Semeval-2025)	p.833
Level 1 1.34	Workshop 18: The Second Workshop on Analogical Abstraction in Cognition, Perception, and Language (Analogy-ANGLE II)	p.834
Level 2 2.15	Workshop 19: BioNLP 2025 and Shared Tasks (BioNLP-ST 2025)	p.835
Level 2 Hall C	Workshop 20: The 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP 2025)	p.836
Level 1 1.31-32	Workshop 21: The First Workshop on Large Language Model Memorization (L2M2)	p.837
Level 2 Hall B	Workshop 22: First Workshop on LLM Security (LLMSEC)	p.838
Level 2 2.44	Workshop 23: MAGMaR: Multimodal Augmented Generation via MultimodAl Retrieval	p.839
Level 1 1.14	Workshop 24: FieldMatters–SIGTYP	p.840
Level 1 Hall N1	Workshop 25: The 3rd workshop on Towards Knowledgeable Foundation Models	p.841
Level 2 2.31	Workshop 26: WikiNLP: Advancing Natural Language Processing for Wikipedia	p.842
Level 2 Hall 2.17	Workshop 27: The 9th Workshop on Online Abuse and Harms (WOAH 2025)	p.843
Level 1 1.61-62	Workshop 28: The First Joint Workshop on Large Language Models and Structure Modeling	p.844

Workshop 01

W01: 6th AfricaNLP Workshop: Multilingual and Multicultural-aware LLMs

David Adelani, Constantine Lignos, Tajuddeen Gwadabe, Bonaventure F. P. Dosso, Israel Abebe Azime, Everlyn Asiko Chimoto, Clemencia Siro, Henok Biadglign Ademtew, Happy Buzaaba, Hatem Haddad, Menno Van Zaanen, Mmasibidi Setaka, Andiswa Bukula, Nomsa Skosana

Thursday, July 31
Level 1 Hall 1.34

Description: The AfricaNLP workshop brings together a diverse group of researchers to explore solutions, collaborations, and innovation around enhancing LLMs' capabilities in African languages and ensuring cultural awareness in their applications.

Workshop 02

W02: The 12th Workshop on Argument Mining

Elena Chistova, Philipp Cimiano, Shohreh Haddadan,
Gabriella Lapesa, Ramon Ruiz-Dolz

Thursday, July 31
Level 2 Hall B

Description: Argument Mining (also known as "argumentation mining") is an emerging research area within computational linguistics that started with focusing on automatically identifying and classifying argument elements, covering several text genres such as legal documents, news articles, online debates, scholarly data, and many more. In recent years, the field (broadly Computational Argumentation) has grown to explore argument quality and synthesis on many levels. The field offers practical uses such as argument-focused search and debating technologies, e.g., IBM Project Debater. The growing interest in computational argumentation has led to several tutorials at major NLP conferences.
Besides providing a forum to discuss and exchange cutting edge research in this field, a secondary goal of this year's edition will be to broaden the disciplinary scope of the workshop by inviting other disciplines (e.g., (computational) social and political science, psychology, humanities) as well as other subareas of NLP to actively participate in the workshop and further shaping the field of argument mining. In particular, we would like to create synergies between the fields of argument mining and natural language reasoning.

Workshop 03

W03: 2nd Workshop on Natural Language Processing meets Climate Change

Dominik Stammbach, Tobias Schimanski, Jingwei Ni, Alba Sun, Alok Singh, Christopher Manning, Gaku Morio, Kalyan Dutia, Peter Henderson, Saeid Vaghefi, Veruska Muccione

Thursday, July 31
Level 2 Hall 2.17

Description: This workshop is dedicated to discuss how Natural Language Processing can be incorporated in Climate Change science, and help mitigating or adapting to climate change.

Workshop 04

W04: Eighth Workshop on Fact Extraction and VERification (FEVER)

Andreas Vlachos, Mubashara Akhtar, Michael Schlichtkrull, Christos Christodoulopoulos, Yulong Chen, Zhenyun Deng, Marek Strong, Arpit Mittal, Oana Cocarascu, Chenxi Whitehouse, James Thorne, Zhijiang Guo, Rami Aly, Rui Cao

Thursday, July 31
Level 2 Hall 2.31

Description: The Fact Extraction and Verification Workshop brings together researchers working on this topic as well as related ones such as recognizing textual entailment, question answering and argumentation mining.

Workshop 05

W05: Workshop on Meaningful, Efficient, and Robust Evaluation of LLMs and GEM: Natural Language Generation, Evaluation, and Metrics

Sebastian Gehrmann, Gabriel Stanovsky, Enrico Santus, Itay Itzhak, João Sedoc, Kaustubh Dhole, Michal Shmueli Scheuer, Miruna Clinciu, Ofir Arviv, Rotem Dror, Simon Mille, Yotam Perlitz, Oyvind Tafjord

Thursday, July 31
Level 2 Hall C

Description: Evaluating large language models (LLMs) is challenging. Running LLMs over medium or large scale corpus can be prohibitively expensive; they are consistently shown to be highly sensitive to prompt phrasing, and it is hard to formulate metrics which differentiate and rank different LLMs in a meaningful way. Consequently, the validity of the results obtained over popular benchmarks such as HELM or MMLU, lead to brittle conclusions (Sclar er al., 2024, Mizrahi et al., 2024, Alzahrani et al., 2024). We believe that meaningful, efficient, and robust evaluation is one of the cornerstones of the scientific method, and that achieving it should be a community-wide goal.
In this workshop we seek innovative research relating to the evaluation of LLMs and language generation systems in general. This includes, but is not limited to, robust, reproducible and efficient evaluation metrics, as well as new approaches for collecting evaluation data which can help in better differentiating between different systems and understanding their current bottlenecks.

Workshop 06

W06: The 19th Linguistic Annotation Workshop (LAW XIX)

Siyao Peng, Ines Rehbein, Amir Zeldes

Thursday, July 31
Level 1 Hall 1.15-16

Description: The Linguistic Annotation Workshop (LAW 2025) targets innovative research on linguistic annotation, focusing on subjectivity and the annotation of concepts from social sciences.

Workshop 07

W07: The 4th Workshop on NLP for Positive Impact

Ruyuan Wan, Katherine Atwell, Laura Biester, Angana Borah, Daryna Dementieva, Oana Ignat, Neema Kotonya, Ziyi Liu, Steven Wilson, Jieyu Zhao

Thursday, July 31
Level 1 Hall N1

Description: The 4th Workshop on NLP for Positive Impact aims to advance innovative NLP research for positive societal impact, emphasizing responsible methods and diverse applications.

Workshop 08

W08: REALM: First Workshop for Research on Agent Language Models

Ehsan Kamalloo, Nicolas Gontier, Xing Han Lu, Shikhar Murty, Alexandre Lacoste, Nouha Dziri, Hanna Hajishirzi, Graham Neubig

Thursday, July 31
Level 1 Hall 1.61-62

Description: The REALM workshop aims to bring together researchers, practitioners, and thought leaders to discuss and align on the current landscape, key challenges, and future directions of LLM Agents.

Workshop 09

W09: SDP 2025: The 5th Workshop on Scholarly Document Processing

Tirthankar Ghosal, Philipp Mayr, Aakanksha Naik, Amanpreet Singh, Anita de Waard, Dayne Freitag, Georg Rehm, Sonja Schimmler, Dan Li

Thursday, July 31
Level 1 Hall 1.14

Description: The goal of this workshop is to provide a venue for addressing these challenges, as well as a platform for tasks and resources supporting the processing of scientific documents. Our long-term objective is to establish scholarly and scientific texts as an essential domain for NLP research, to supplement current efforts on web text and news articles.

Workshop 10

W10: The Third Workshop on Social Influence in Conversations (SICon 2025)

James Hale, Deuksin Kwon, Kushal Chawla, Ritam Dutt, Muskan Garg, Liang Qiu, Alexandros Papangelis, Gale Lucas, Zhou Yu, Daniel Hershcovich

Thursday, July 31
Level 1 Hall 1.31-32

Description: Social influence (SI) is the change in an individual's thoughts, feelings, attitudes, or behaviors from interacting with another individual or a group. For example, a buyer uses SI skills to negotiate trade-offs and build rapport with the seller. SI is ubiquitous in everyday life, and hence, realistic human-machine conversations must reflect these dynamics, making it essential to model and understand SI in dialogue research systematically. This would improve SI systems' ability to understand users' utterances, tailor communication strategies, personalize responses, and actively lead conversations. These challenges draw on perspectives not only from NLP and AI research but also from Game Theory, Affective Computing, Communication, and Social Psychology.

Workshop 11

W11: Slavic NLP-2025: The 10th Biennial Workshop on Slavic NLP

Jakub Piskorski, Michał Marcińczuk, Preslav Nakov, Nikola Ljubešić,
Pavel Přibáň, Roman Yangarber

Thursday, July 31
Level 2 Hall 2.44

Description: Slavic NLP 2025 is a Workshop that addresses Natural Language Processing (NLP) for the Slavic languages.
The goal of this Workshop is to bring together researchers from academia and industry working on NLP for Slavic languages. In particular, the Workshop will serve to stimulate research and foster the creation of tools and resources for these languages. We also organize the shared task on the Detection and Classification of Persuasion Techniques in Texts for Slavic Languages.

Workshop 12

W12: The 4th Table Representation Learning Workshop at ACL 2025

Madelon Hulsebos, Shuaichen Chang, Wenhu Chen, Filip Gralinski,
Qian Liu, Huan Sun

Thursday, July 31
Level 2 Hall 2.15

Description: "Tables are a promising modality for representation learning and generative models with too much application potential to ignore. The Table Representation Learning (TRL) workshop is the premier venue in this emerging research area and has three main goals: (1) Motivate structured data (e.g. tables) as a primary modality for representation and generative models and advance the area further. (2) Showcase impactful applications of pretrained table models and identify open challenges for future research, with a particular focus on progress in NLP for this edition at ACL in 2025. (3) Foster discussion and collaboration across the NLP, ML, IR and DB communities."

Workshop 13

W13: The Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)

Mariana Romanyshyn, Olena Nahorna, Oleksii Ignatenko, Andrii Hlybovets

Thursday, July 31 and Friday, August 1, 2025
Virtually on Gather

Description: The UNLP workshop is dedicated to the development of language resources, tools, and NLP solutions for the Ukrainian language. This event brings together professionals from academia and industry who work with Ukrainian or do cross-lingual research that can be applied to Ukrainian. Every year UNLP hosts a shared task and invited speakers who promote research on low-to-mid-resource languages.

Workshop 14

W14: The 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)

Ekaterina Kochmar, Marie Bexte, Andrea Horbach, Ronja Laarmann-Quante, Anaïs Tack, Bashar Alhafni, Victoria Yaneva, Jill Burstein, Zheng Yuan

Thursday, July 31 and Friday, August 1, 2025
Level 1 Hall 1.85-86

Description: The BEA Workshop is a leading venue for NLP innovation in the context of educational applications.

Workshop 15

W15: 29th Conference on Computational Natural Language Learning (CoNLL)

Gemma Boleda, Michael Roth, Emily Cheng, Selina Meyer, Snigdha Chaturvedi

Thursday, July 31 and Friday, August 1, 2025
Level 1 Hall M1

Description: CoNLL is a yearly conference organized by SIGNLL (ACL's Special Interest Group on Natural Language Learning), focusing on theoretically, cognitively, and scientifically motivated approaches to computational linguistics. In particular, it seeks to explore the interaction between theoretical issues in linguistics and cognition, on the one hand, and computational modeling, on the other.

Workshop 16

W16: The 22nd International Conference on Spoken Language Translation (IWSLT 2025)

Marcello Federico, Alex Waibel, Elizabeth Salesky, Jan Niehues, Sebastian Stüker, Atul Kr. Ojha, Marine Carpuat

Thursday, July 31 and Friday, August 1, 2025
Level 1 Hall N2

Description: The International Conference on Spoken Language Translation (IWSLT) is an annual scientific conference dedicated to all aspects of spoken language translation. For more than 20 years, the conference has published research papers and organized key evaluation campaigns within the field, including the creation of key data suites, benchmarks, metrics and new tasks. IWSLT is the annual meeting of SIGSLT, the ACL-ISCA-ELRA Special Interest Group on Spoken Language Translation.

Workshop 17

W17: The 19th Workshop on Semantic Evaluation (Semeval-2025)

Sara Rosenthal, Aiala Rosá, Marcos Zampieri, Debanjan Ghosh

Thursday, July 31 and Friday, August 1, 2025
Level 1 Hall M2

Description: The Semantic Evaluation (SemEval) workshops focus on the evaluation and comparison of systems that analyze diverse semantic phenomena in text, with the aim of extending the current state of the art in semantic analysis and creating high quality annotated datasets in a range of increasingly challenging problems in natural language semantics.

Workshop 18

W18: The Second Workshop on Analogical Abstraction in Cognition, Perception, and Language (Analogy-ANGLE II)

Filip Ilievski, Giulia Rambelli, Marianna Bolognesi, Ute Schmid, Pia Sommerauer

Friday, August 1, 2025
Level 1 Hall 1.34

Description:
The Second Workshop on Analogical Abstraction in Cognition, Perception, and Language (Analogy-Angle II). Explore, model, and understand analogical reasoning in cognition, language, and computational models from an interdisciplinary perspective Analogy-Angle II is a multidisciplinary workshop to advance research on analogical abstraction by bridging the fields of computational linguistics, artificial intelligence, and cognitive psychology. This workshop seeks to foster collaboration among researchers by providing a platform for sharing novel insights, benchmarks, methodologies, and analogy applications across disciplines.

Workshop 19

W19: BioNLP 2025 and Shared Tasks (BioNLP-ST 2025)

Dina Demner-Fushman, Sophia Ananiadou, Makoto Miwa, Jun-ichi Tsujii

Friday, August 1, 2025
Level 2 Hall 2.15

Description: BioNLP associated with ACL SIGBIOMED is an established primary venue for presenting research in language processing for the biomedical domains. The workshop has been running every year since 2002 and continues getting stronger. BioNLP truly encompasses the breadth of the domain and brings together researchers in biomedical and clinical NLP from all over the world.

Workshop 20

W20: The 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP 2025)

Christine Basta, Marta R. Costa-jussà, Agnieszka Falénska,
Debora Nozza, Karolina Stańczak

Friday, August 1, 2025
Level 2 Hall C

Description: The GeBNLP workshop provides a dedicated forum to address gender and other demographic biases in natural language processing models. It aims to foster awareness, share research on mitigating bias through data and algorithmic approaches, and build community consensus around standard evaluation tasks and metrics.

Workshop 21

W21: The First Workshop on Large Language Model Memorization (L2M2)

Robin Jia, Eric Wallace, Yangsibo Huang, Tiago Pimentel, Pratyush Maini, Verna Dankers, Johnny Wei, Pietro Lesci

Friday, August 1, 2025
Level 1 Hall 1.31-32

Description: The First Workshop on Large Language Model Memorization (L2M2), co-located with ACL 2025 in Vienna, seeks to provide a central venue for researchers studying LLM memorization from these different angles.

Workshop 22

W22: First Workshop on LLM Security (LLMSEC)

Leon Derczynski, Jekaterina Novikova, Muhao Chen

Friday, August 1, 2025
Level 2 Hall B

Description: Work on adversarially-induced failure modes of large language models, the conditions that lead to them, and their mitigations.

Workshop 23

W23: MAGMaR: Multimodal Augmented Generation via MultimodAl Retrieval

Reno Kriz, Kenton Murray, Kate Sanders, Eugene Yang, Cameron Carpenter, Francis Ferraro, Benjamin Van Durme

Friday, August 1, 2025
Level 2 Hall 2.44

Description: Vast amounts of information today is being stored as videos with minimal text metadata, necessitating further research around the efficient discovery, understanding, and synthesis of multimodal collections. To address this need, the workshop on Multimodal Augmented Generation via Multimodal Retrieval (MAGMaR) workshop will focus on two primary areas: (1) the retrieval of multimodal content, which spans text, images, audio, video, and multimodal data (e.g., image-language, video-language); and (2) retrieval-augmented generation, with an emphasis on multimodal retrieval and generation.

Workshop 24

W24: FieldMatters–SIGTYP: The 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP) and the Fourth Workshop on NLP Applications to Field Linguistics (FieldMatters)

Michael Hahn, Ekaterina Vylomova

Friday, August 1, 2025
Level 1 Hall N1

Description: SIGTYP is the first dedicated venue for typology-related research and its integration in multilingual NLP.

Workshop 25

W25: The 3rd workshop on Towards Knowledgeable Foundation Models

Yuji Zhang, Xiaozhi Wang, Mor Geva, Chi Han, Shangbin Feng, Silin Gao, Sha Li, Manling Li, Heng Ji

Friday, August 1, 2025
Level 1 Hall 1.14

Description: Knowledge has been an important pre-requisite for a variety of AI applications, and is typically sourced from either structured knowledge sources such as knowledge bases and dictionaries or unstructured knowledge sources such as Wikipedia documents. More recently, researchers have discovered that language models already possess a significant amount of knowledge through pre-training: LLMs can be used to generate commonsense knowledge and factual knowledge context for question answering. While the results are encouraging, there are still lingering questions: Where does this knowledge come from? How much do language models know? Is this knowledge reliable? If some knowledge is wrong, can we fix it?

Workshop 26

W26: WikiNLP: Advancing Natural Language Processing for Wikipedia

Akhil Arora, Isaac Johnson, Lucie-Aimée Kaffee, Tzu-Sheng Kuo,
Tiziano Piccardi, Indira Sen

Friday, August 1, 2025
Level 2 Hall 2.31

Description: WikiNLP invites researchers to contribute novel uses of Wikimedia data or studies of the impact of Wikimedia data within the NLP community. We will discuss tensions around multilinguality and concerns raised by generative AI. We will also highlight successful approaches to developing tooling that can assist the Wikimedia community in maintaining and improving the breadth of the Wikimedia projects.

Workshop 27

W27: The 9th Workshop on Online Abuse and Harms (WOAH 2025)

Zeerak Talat, Agostina Calabrese, Christine de Kock, Francielle Vargas, Flor Miriam Plaza del Acro, Debora Nozza

Friday, August 1, 2025
Level 2 Hall 2.17

Description: WOAH, the 9th Workshop on Online Abuse and Harms, invites paper submissions from a diverse range of fields. Digital technologies have brought significant benefits to society, transforming how people connect, communicate, and interact. However, these same technologies have also enabled the widespread dissemination and amplification of abusive and harmful content, such as hate speech, harassment, and misinformation. Given the sheer volume of content shared online, addressing abuse and harm at scale requires the use of computational tools. Yet, detecting and moderating online abuse remains a complex task, fraught with technical, social, legal, and ethical challenges.

Workshop 28

W28: The First Joint Workshop on Large Language Models and Structure Modeling

Zeerak Talat, Agostina Calabrese, Christine de Kock, Francielle Vargas, Flor Miriam Plaza del Acro, Debora Nozza

Friday, August 1, 2025
Level 1 Hall 1.61-62

Description: The 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025) at ACL 2025 aims to encourage discussions and highlight methods for language structure modeling in the era of LLMs.

13

Poster Printing

XEST is the ACL 2025 preferred **poster printing** company. The daily **deadline** to order for next-day pick-up is 11 a.m.

Poster pick-up times will be daily, 08:00 - 15:00, Monday, July 28 - Friday, August 1 on Level 0, Hall 0.11



Last mintue order? Scan the QR code below.
The daily deadline to order for next-day pick-up is 11 a.m.



Alternative near Austria Center Vienna: https://www.iprinter.at/copyshop.html
Alternative in the City Center: https://diekopie.at/poster-plane-plakate/

14 Local Guide

About Vienna

Architecturally, the city is still characterized primarily by the buildings around Vienna's Ringstrasse from the Wilhelminian period, but also by Baroque and Art Nouveau (Vienna Modernism and Vienna Secession). Through its role as the imperial capital and residence of the Austrian Empire from 1804 onwards, Vienna became a cultural and political center of Europe. The historic center of Vienna and Schönbrunn Palace are now UNESCO World Heritage sites. With around 7.5 million tourists annually and around 18.9 million overnight stays in 2024, Vienna is one of the most visited cities in Europe.

Language

German is considered the official language of Austria, however English is widely spoken throughout the country.

Safety

Vienna is generally considered a very safe place to visit, but like visiting any major city, be aware of your surroundings and of scams, know emergency numbers (listed below), keep digital copies of important documents (passport).

Power Plugs/Outlets

Outlets in Vienna are European Type F and the voltage is 230 volts, 50 Hz. This means you'll need a travel adapter or possible converter for devices from countries like the US, which have different plug types and voltages.

In case of an Emergency

European Emergency Number: 112 (Ambulance, Fire, Police)

Clothing

Plan on packing light, breathable summer clothing. Please note that in Vienna there can be very hot summer days and heavy rain.Water mists are provided at certain locations for cooling refreshment. The City of Vienna also provides free water from drinking fountains.

Currency & Living Costs

The currency used is the 'Euro' written as EUR or €.

Euro notes start from 5 € and go up to 200 €, whereas the coins have denominations of 1 cent, 2 cents, 5 cents, 10 cents, 20 cents, 50 cents, 1 €, and 2 €.

Average Cost of some items

- Business Lunch 25 €
- Cup of Coffee 4 €
- Filled Sandwich 5 €
- Mineral Water 2 €

Most stores accept credit and debit cards

Water

The tap water in Vienna has a very high quality and is safe to drink.

Vienna Sights and Tourism Info Desk

The Vienna Tourism Board will provide a tourism info desk directly at the conference venue from July 28 to 30, 2025 for your convenience. The provided services include:

- Tourist information about Vienna

- Event details and recommendations
- Assistance with hotel bookings
- Various brochures about Vienna, including city maps in multiple languages

Feel free to also explore all attractive sights and events in Vienna around the time of the conference here.
The official Vienna City Card provides you with discounts for public transportation and tourist attractions. Visit `https://www.viennacitycard.at/en/` for more information.

Popular Destinations

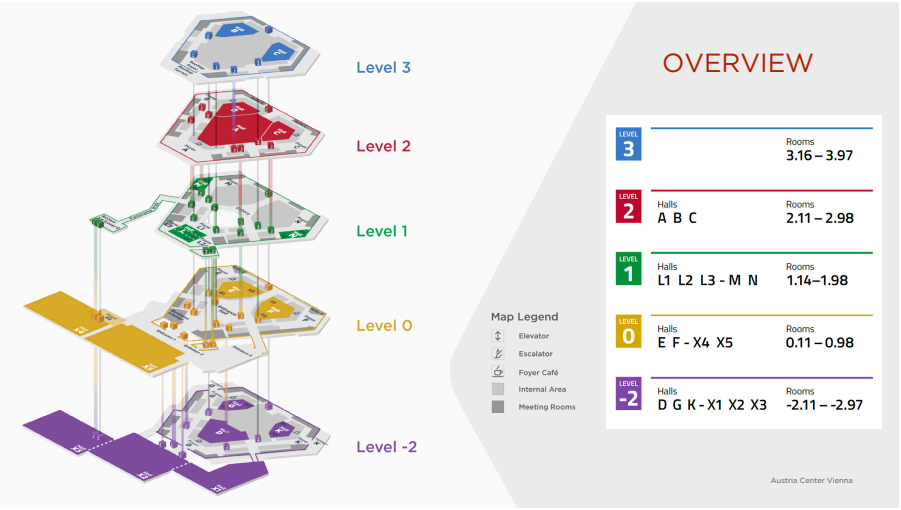
- Albertina (Art museum: Monet, Picasso and others, `https://www.albertina.at/en/home`)
- Kunsthistorisches Museum (Museum of Art History, very large collection: Tizian, Raffael, Caravaggio, Dürer, Bruegel, Rubens, Vermeer, Rembrandt, ...) `https://www.khm.at/`
- Riesenrad (Giant ferris wheel in the Prater amusement park) `https://wienerriesenrad.com/en/home-2`
- Stephansdom (Cathedral, `https://www.stephanskirche.at/visit.php`)
- Schloss Schönbrunn (Castle, `https://www.schoenbrunn.at/en/`)
- Schloss Belvedere (Castle and art museum, holds Klimt's famous work "The Kiss" `https://www.belvedere.at/en`)

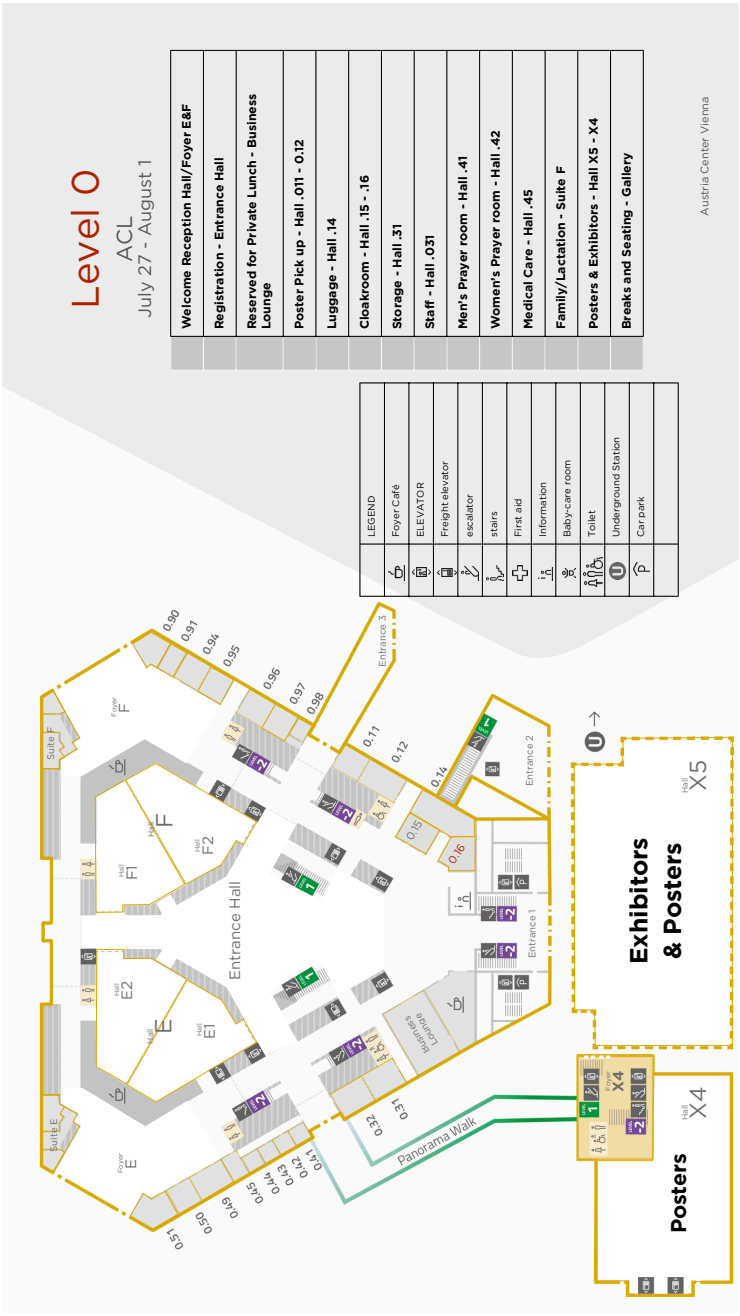
Popular Shopping Areas

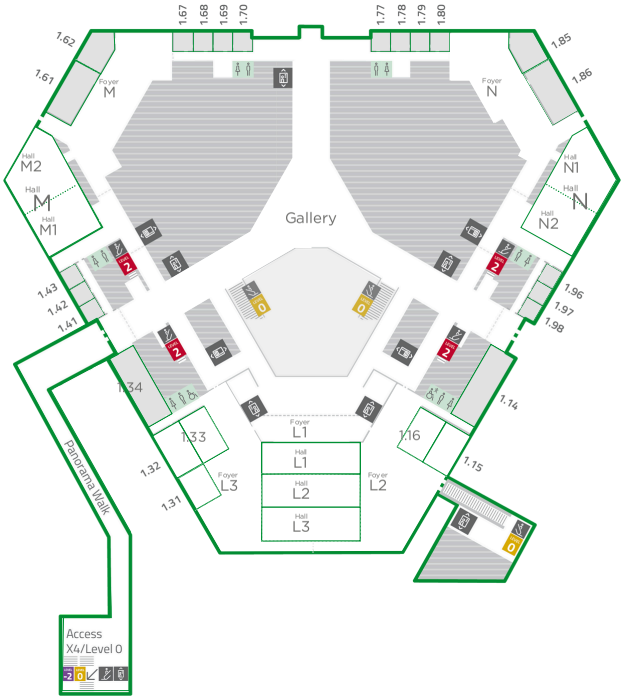
- Mariahilfer Straße: `https://www.wien.info/de/sehen-erleben/shopping/shopping-rund-um-die-altstadt`
- Kärntner Straße: `https://www.wien.info/en/see-do/shopping/old-city`

15

Venue Map


Level 3
Level 2
Level 1
Level 0
Level -2
OVERVIEW
LEVEL 3
Rooms 3.16 – 3.97
LEVEL 2
Halls A B C
Rooms 2.11 – 2.98
LEVEL 1
Halls L1 L2 L3 - M N
Rooms 1.14–1.98
LEVEL 0
Halls E F - X4 X5
Rooms 0.11 – 0.98
LEVEL -2
Halls D G K - X1 X2 X3
Rooms -2.11 – -2.97
Map Legend
Elevator
Escalator
Foyer Café
Internal Area
Meeting Rooms
Austria Center Vienna


Level 0
ACL
July 27 - August 1
Welcome Reception Hall/Foyer E&F
Registration - Entrance Hall
Reserved for Private Lunch - Business Lounge
Poster Pick up - Hall .011 - 0.12
Luggage - Hall .14
Cloakroom - Hall .15 - .16
Storage - Hall .31
Staff - Hall .031
Men's Prayer room - Hall .41
Women's Prayer room - Hall .42
Medical Care - Hall .45
Family/Lactation - Suite F
Posters & Exhibitors - Hall X5 - X4
Breaks and Seating - Gallery
LEGEND
Foyer Café
ELEVATOR
Freight elevator
escalator
stairs
First aid
information
Baby-care room
Toilet
Underground Station
Car park
Austria Center Vienna
Entrance Hall
Entrance 1
Entrance 2
Entrance 3
Panorama Walk
Exhibitors & Posters
Hall X5
Posters
Hall X4
Foyer E
Foyer F
Suite E
Suite F
Hall E1
Hall E2
Hall F1
Hall F2
Business Lounge



	LEGEND
	ELEVATOR
	Freight elevator
	escalator
	stairs
	Baby-care room
	Toilet

Level 1

ACL
July 27 - August 1

	Tutorials (Sunday) **Halls: M.1-2, N.1-2**
	Orals (Mon-Wed) **Halls: M.1-2, N.1-2, L.2-3, 1.15-16, 1.61, 1.62, 1.85, 1.86**
	Workshops (Thur & Fri) **Halls: M.1-2, N.1-2, L.1-3** **1.14-16, 1.31-32, 1.34, 1.61, 1.62, 1.85, 1.86**
	Workshop Posters (Thur & Fri) **Hall L.1- L.3**
	Birds of a Feather **Halls: 1.14, 1.31-32, 133**

Austria Center Vienna


Level 2
ACL
July 27 - August 1
Tutorials (Sunday) Halls B & C
Plenary (Mon-Wed) Halls A-C
Panels/Orals (Wed) Halls A-C
Workshops (Thur & Fri)
Halls: B &C, 2.15, 2.17, 2.31, 2.44
Adhoc - Hall 2.12
LEGEND
Foyer Café
ELEVATOR
Freight elevator
escalator
stairs
Baby-care room
Access for Wheelchairs
Toilet
Austria Center Vienna
Hall A
Hall B
Hall C
Foyer A
Foyer B
Foyer C
Suite B
Suite C
2.11
2.12
2.13
2.14
2.15
2.17
2.31
2.32
2.33
2.34
2.41
2.42
2.43
2.44
2.61
2.62
2.82
2.83
2.95
2.96
2.97
2.98

Alibaba Cloud Careers

Alibaba Cloud Campus Recruitment

About Alibaba Cloud

Apsara Lab - Data to Intelligence Lab

Alibaba Cloud's Data to Intelligence Lab, led by world-renowned AI expert Prof. YE Jieping, is driving intelligent transformation across society with Qwen models. The team focuses on industry-specific LLMs, multimodal models, and foundational research on large models. Their innovations have been applied in various sectors, including smart traffic systems, urban governance, and biomedicine, while also enhancing major global sports events like the Asian Games and the Olympics.

Alibaba Cloud Computing Platform

Alibaba Cloud Computing Platform includes a proprietary big data platform ODPS (MaxCompute/Hologres/DataWorks, etc.), open-source big data platforms (E-MapReduce/Realtime Compute for Apache Flink, etc.), and an artificial intelligence platform PAI, etc. It provides a comprehensive product system covering data collection, storage and analysis, development and governance, AI training and application. Meanwhile, Alibaba Cloud Computing Platform is also equipped with a variety of computing capabilities including large-scale batch processing, real-time streaming processing, and machine learning. It can concurrently schedule over 100,000 servers for parallel computing. Alibaba Cloud is the only Chinese Company in Forrester's Leaders quadrant for cloud data warehouses. In terms of AI infrastructure, related products of Alibaba Cloud Computing Platform ranked second globally and first in Asia for product and technical capabilities in the Forrester Wave: AI Infrastructure Solutions 2024 evaluation.

</SHAPE YOUR OWN FUTURE_

Application method

Scan the QR code to submit your resume

Target students

Currently enrolled university students

Positions

Apsara Lab–Data to Intelligence Lab

- LLM Algorithm Research Intern

Alibaba Cloud Computing Platform

- Research on the application of large model downstream task transfer technology in general content security and agent security
- Research on knowledge distillation techniques for LLMs
- Research on key technologies for the pre-training of ultra-large-scale MoE models
- Exploration of agent-based adaptive RAG
- Optimization of large model hallucination in RAG scenarios

Locations: Hangzhou, Beijing

</SHAPE YOUR OWN FUTURE_

ANT GROUP JOBS



ANT STAR
Plan A Talent Program

Join Ant Group and become a member of "Plan A"
Light up the future with technology together!

"Plan A" is a new elite AI talent initiative under "Ant Star." **It aims to recruit world-class artificial intelligence researchers from leading global universities to join Ant Group's exploration of AGI.**

The letter "A" embodies two core ideals: Apex, Visionary pioneers reaching the summit, daring to push the boundaries of intelligence. Action, Pragmatic innovators committed to ensure AI benefits all humanity.

We discover and empower exceptional rising stars worldwide– passionate tech innovators with outstanding research enthusiasm.Through top-tier challenges, resources, and development spaces,we prepare tomorrow's AI pioneers.

Scan the code to learn about Ant Group

The Minds We Seek

Top talents in LLM, Multimodal, Infra, RL, Diffusion, Sparse Attention, and Synthetic data.

Recruitment Position:

Research Scientist -

Reinforcement Learning | Deep Learning | Synthetic Data | Diffusion | Inference | LLM Pre-training | LLM Post-training | Multimodal

What We Offer

Frontier Technology Arena
- Industry pioneer mentorship
- Mission-critical R&D challenges
- Strategic core tech domains

Unconstrained Innovation
- Personalized development roadmaps
- Full research freedom
- Unlimited GPU resources

Elite Growth Ecosystem
- Levels and salaries based on ability, not seniority
- Get recognized faster for what you achieve

Global R&D network
- Hubs across Hangzhou/ Beijing/Shanghai/Singapore
- Integrated research labs
- Cross-disciplinary exchanges

Recruitment Target

Eligibility: Open to master's/PhD graduates in computer science, software engineering, AI, cybersecurity, ICT, mathematics, statistics, and interdisciplinary STEM fields.

Graduation Window: Candidates graduating in or after November 2025.

Join Us

Scan the QR code to complete the resume submission.

Contact Us:
AntTalent@antgroup.com

AI at Bloomberg

Bloomberg is building the world's most trusted information network for financial professionals. Our 9,000+ engineers, developers, and data scientists are dedicated to advancing and building new solutions and systems for the Bloomberg Terminal and other products in order to solve complex, real-world problems.

We have 15+ years of experience building and using AI solutions to help process and organize the ever-increasing volume of structured and unstructured financial information. In fact, we were one of the first on the street to use AI. Our team of 350+ AI researchers and engineers deploy different AI technologies, including ML, NLP, computer vision, information retrieval, and generative models, as well as neural networks and time-series analysis. This is how we provide our customers around the globe with the reliable, high-quality data, news, and information they need to make well-informed decisions and be more productive.



Learn more here:

#MAKEITHAPPENHERE

Make it happen here.

Bloomberg

Engineering

CITADEL | CITADEL | Securities

Build, test, and refine your ideas at the speed of the markets.

The markets present rich data, complex problems and fast feedback. We treat market problems as scientific questions. Our researchers can quickly see how the markets validate or refute their thinking.

Our culture of achievement invites the brightest minds in science. As a result, we have some of the most talented and accomplished researchers in the world working alongside our investment professionals, traders, and engineers.



Learn more at
events.citadel.com/acl2025

citadel.com | citadelsecurities.com



cohere

The all-in-one platform for private and secure AI.

@cohere
cohere.com
cohere

Cohere Labs

Exploring the unknown, together. Changing spaces where breakthroughs happen.

@cohere_labs
cohere.com/research
cohere labs



Where your skills meet the right opportunity

LinkedIn

About Xiaomi

Xiaomi Corporation was founded in April 2010 and listed on the Main Board of the Hong Kong Stock Exchange on July 9, 2018 (1810.HK). Xiaomi is a consumer electronics and smart manufacturing company with smartphones and smart hardware connected by an IoT platform at its core.

Embracing our vision of "Make friends with users and be the coolest company in the users' hearts", Xiaomi continuously pursues innovations, high-quality user experience and operational efficiency. The company relentlessly builds amazing products with honest prices to let everyone in the world enjoy a better life through innovative technology.

Xiaomi Top Campus Talent Recruitment

100+cutting-edge research topics, orchestrating the future of technological innovation!

Covering frontier domains: Large Models, Multimodality, Machine Learning, Embodied Intelligence, Semiconductor Chips, Autonomous Driving, etc.

- Elite Mentorship from Top Industry Experts
- Premium Resource Allocation for Cutting-Edge Research
- Boundless Career Growth Opportunities
- Engagement in Pioneering Research Projects
- Competitive Compensation & Benefits

Recruitment Targets

Internship Program:

Current undergraduate, master's or PhD students

Campus recruitment:

- Graduates graduating between 2023 and 2026 (inclusive)
- Postdoctoral researchers completing their programs in 2026

Base: Beijing | Shanghai | Wuhan

Join Us



Pushing the boundaries of AI through research, infrastructure and product innovation.

ACL 2025

amazon

Amazon Science

Academics at Amazon

Internships at Amazon

Come build the future with us

At Amazon, we believe that scientific innovation is essential to being the most customer-centric company in the world. Our scientists' ability to have an impact at scale allows us to attract some of the brightest minds in artificial intelligence and related fields.



文心大模型

BAIDU WENXIN

A series of knowledge-enhanced large models, from general-purpose to industry-specific, independently developed by Baidu, speed up AI innovations and empower the industry to upgrade.

ERNIE

2025.3 - 2025.4
ERNIE 4.5 / 4.5 Turbo
- Advanced Language Proficiency
- Multimodality

ERNIE X1 / X1 Turbo
- Deep Thinking
- Function Calling

2024.6
ERNIE 4.0 Turbo
- Knowledge Distillation

2023.10
ERNIE 4.0
- Incremental Learning
- Multi-dimensional Reward Model
- Agent

2023.5
ERNIE 3.5
- Factual Knowledge Enhancement
- Reasoning Enhancement

2023.3
ERNIE Bot
- Knowledge Enhancement
- Retrieval Augmentation
- Dialogue Enhancement

2021.7
ERNIE 3.0
- Knowledge-infused Learning

2019.7
ERNIE 2.0
- Knowledge-guided Multi-task Learning

2019.3
ERNIE 1.0
- Multi-granularity Knowledge Fusion Learning

Knowledge Enhancement	Retrieval Augmentation	Dialogue Enhancement	Agents
Supervised Fine-Tuning	Reinforcement Learning	Prompt	
Pre-Training Models			

ByteDance | Seed

About Seed Team

The ByteDance Seed team was established in 2023, dedicated to discovering new approaches to general intelligence, pushing the boundaries of AI, and exploring new forms of interaction. The team's research areas include LLM, speech, vision, world models, infrastructure, AI Infra, and next-generation AI interactions. The team operates labs and positions in China, Singapore, the United States, and other locations.

With a long-term vision and determination in the AI field, the ByteDance Seed team remains committed to foundational research and aims to become a world-class AI research team, contributing to technological and societal advancement. The team has already launched industry-leading general large models and cutting-edge multimodal capabilities, supporting over 50 application scenarios such as Doubao, Coze, and Jimeng.

Jobs Information

Job Locations

Beijing, Shanghai, Hangzhou, Shenzhen, Singapore, San Jose, Seattle, and more.

Scan the QR code and check out the positions of Seed team

Follow Us

Visit the official website of the Seed team

seed.bytedance.com

Stay updated on our latest research, model performance, and job opportunities by following us on **RED** and **Zhihu**

小红书 Seed 研究员

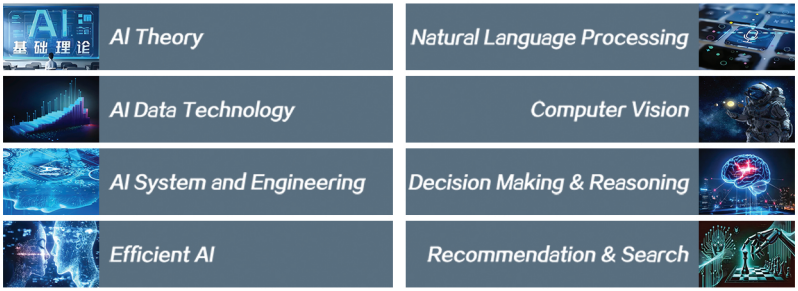
知乎 字节跳动 Seed

Scan the QR code to follow the official WeChat account

Noah's Ark Lab

HUAWEI

Founded in 2012, Noah's Ark Lab is the AI research centre for Huawei Technologies. We undertake the following primary domains: AI theory, AI data technology, AI system and engineering, efficient AI, computer vision, natural language processing, decision making & reasoning, recommendation & search, to drive fundamental R&D in AI.



Mission

Our mission is to build practical AI technology to solve real world problems. We entail cutting-edge AI technologies including reinforcement learning, unsupervised learning, large foundation models, etc., in multiple areas, such as network intelligence (automatic control, security and maintenance), enterprise intelligence (e.g., safe city, smart supply chain, smart accounting) and terminal intelligence (e.g., AI camera, smartphones, and personalized recommendation).

Collaboration

We have top-tier researchers from world-leading universities and industry pioneers, collaborating with global innovative organizations including 40+ universities and 1,000+ researchers from 10+ countries. Through joint labs with 10+ top-tier AI universities, we pioneer breakthroughs in large language models, multimodal AI, agent systems, etc. Our achievements include multiple best/outstanding paper awards at top-tier conferences, building Huawei's competitiveness in AI.

Global Vision

Now, Noah's Ark Lab has established global footprint including Shenzhen, Beijing, Shanghai, Xi'an, Nanjing, Hefei, Hong Kong, Singapore, Paris, London, Montreal, Toronto, etc, and built global research partnership with over 12 top universities.

Join Us

We are hiring researchers in the related areas at all levels, including junior and senior positions. Interested candidates should send application materials including resume to noahlab1@huawei.com.





Inventing What's Next

IBM Research is home to 3,000 scientists and researchers who deeply believe in the power of the scientific method to invent what's next in computing for our company, for our clients and the world. We are pioneering technologies that will transform industries and society, including the future of AI, hybrid cloud, quantum computing and semiconductors. We've been here since the earliest days of computing, and we're leading the charge for the future. Since our first lab opened in 1945, we've authored more than 110,000 research publications. Our researchers have won six Nobel Prizes, six Turing Awards, and IBM has been granted more than 150,000 patents.

ibm.research.com



Megagon Labs

Connecting AI, Data, and People for Better Opportunities

ACL Main:
CypherBench: Towards Precise Retrieval over Full-scale Modern Knowledge Graphs in the LLM Era

ACL Findings:
FactLens: Benchmarking Fine-Grained Fact Verification



blue

Research that ships.

Blue is an open-source agentic framework for building compound AI systems that integrate language models, data, APIs, and agents together for performing complex tasks. It's built for researchers exploring agentic systems, planning, data management, and data centric NLP in enterprise scale settings. Use Blue for your projects or contribute to it's development.
Go to our Github →



ORACLE

Real careers in artificial intelligence

Oracle is proud to sponsor the 63rd Annual Meeting of the Association for Computational Linguistics. And we're just as proud to drive the future of AI, leveraging cutting-edge agentic platforms and multimodal GenAI to empower customers. Our teams are pushing the boundaries of innovation, from reasoning and planning to retrieval-augmented generation, GenAI for Databases and Analytics, NL2Code, and model architecture for large-scale training/inference.

Now, it's your turn. Push your career at the forefront of innovation and put your knowledge to work with the cloud leader for global business.

Take the next step in your career. Scan the QR code to explore opportunities with the Oracle Cloud Infrastructure AI Science team.

Create the **future** with us



WHERE DATA DOES MORE

Snowflake AI Data Cloud.
The platform for the AI era.

snowflake.com

snowflake®

TENCENT PROJECT UP

腾 讯 青 云 计 划



About PROJECT UP

Project UP is a Tencent talent program with the aim of recruiting top technology students from around the world

We Connect - Campus Collaboration
A new collaboration model that enables you to join Tencent's core research projects on campus

We Discover - Global Premier Events
Expanded-scale events: UP Open Days, City Exclusives to Starlit Tech Gala

We Inspire - Expertise Meets Excellence
Engagement with Tencent's tech executives and distinguished scientists

Meet Us at Booth 12

Tencent Project UP Official Website

Tencent Project UP Global Information

Connect with our HR Team

Toloka

Expertly crafted data for AI agents & models

1. Demonstrations generation for SFT
2. Preferences collection for RLHF / DPO
3. Evaluation and Red Teaming

- All data types (text, image, video, audio)
- Diverse languages
- Specialized domains (coding, law, finance, more)

toloka.ai

LLM Research on the world's most exciting data: Machine Learning & NLP at Thomson Reuters

Train, evaluate, create data for, prototype with, and adversarially test models trained on data created by the world's most experienced journalists, editors, and lawyers. Research and experiment in collaboration with world-leading universities and publish at the best venues for LLM Research.

Join our team

© 2025 Thomson Reuters TR5806232/6-25

Thomson Reuters™

KUAISHOU



About Kuaishou

Kuaishou is a leading global content community and social platform. Kuaishou has developed several cutting-edge models such as the Large Language Model "KwaiYii", Video Generation Large Model "Kling", Image Generation Large Model "Kolors", and Large Language Model for Coding "Kwaipilot". These models are deeply integrated into Kuaishou's core businesses, including short videos, live streaming, online advertising, e-commerce, and local services. They have also inspired innovative applications like "KlingAI", "AI Xiao Kuai", and "Kwaipilot".

Join Us!

The KSTAR-X Talents Program is hiring now.
Scan the QR code to learn more and submit your application, or reach out via campus@kuaishou.com



Appen

Build Safer, Smarter AI with High-Quality, Bespoke Data

For over 25 years, Appen has enabled the world's top AI teams to fine-tune and evaluate models using high-quality, human-labeled data across 500+ languages and 200+ countries.

From multilingual LLMs to complex multimodal pipelines, our experts deliver the precision your models need.

📍 Visit us at Booth 8 🔗 appen.com



Adobe supports ACL 2025

Adobe Research is Hiring!

Our Research Areas

- Artificial Intelligence & Machine Learning
- Computer Vision, Imaging & Video
- Graphics (2D & 3D)
- Content Intelligence
- Audio
- Document Intelligence
- AR, VR & 360 Photography
- Natural Language Processing
- Human-Computer Interaction
- Data Intelligence
- Intelligent Agents & Assistants
- Systems & Languages

If you're interested in research opportunities at Adobe scan the QR code to submit your contact information and resume.

Learn more:
Interns: Research Adobe Express Page
Experience: research.adobe.com

Register here or scan the QR code

Babelscape®

Language generation, semantically grounded

Explore our technologies in detail
just scan the QR code below

https://babelscape.com

DEFINED.AI

Empower your **LLM**

Discover powerful, responsible, and high-quality AI.
Smarter fine-tuning.
Safer data.
Superior models.

Find out more

JPMorganChase

MACHINE LEARNING
CENTER OF EXCELLENCE

Meet the MLCOE, leaders in machine learning and innovation. Visit us to learn how our diverse team is delivering AI/ML solutions to transform finance!

We're looking for curious problem-solvers with a passion for developing innovative ML solutions.

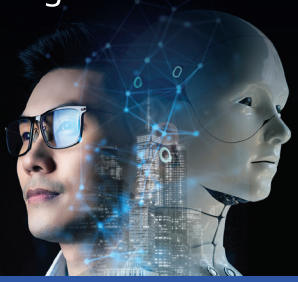
jpmorgan.com/mlcoe

See you at
Booth 5!

NEC

\Orchestrating a brighter world

Advancing agentic
AI for greater social value.



#NECLabs
neclab.eu

translated.

$100,000

to fund language technology innovators who share the goal of making it easier for everyone to understand and be understood by all others.

Find out more.

DataoceanAI

Empower Your AI with Best Data

OFF-THE-SHELE DATASETS

DATA SERVICES

DOTS PLATFORM

MODEL EVALUATION

CONTACT US

Dataocean AI Inc.
100 N Howard St Ste R, Spokane, WA 99201
contact@dataoceanai.com
www.dataoceanai.com

Thank You to Our Sponsors

DIAMOND

CITADEL | CITADEL Securities

ANT GROUP

Bloomberg

cohere

Google Research

LinkedIn

Meta

Alibaba Cloud

PLATINUM

ORACLE

Toloka

ByteDance

IBM

HUAWEI

Megagon Labs

Baidu 百度

amazon | science

Tencent

snowflake

GOLD

Thomson Reuters

CHEN TIANQIAO & CHRISSY INSTITUTE

KUAISHOU

SILVER

Appen

Adobe

nexdata

BRONZE

Babelscape

JPMorganChase

NatWest Group

DEFINED.AI

NEC

DataoceanAI

translated.