

BlackboxNLP 2026

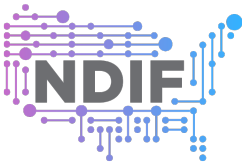
**The 9th BlackboxNLP Workshop: Analyzing and
Interpreting Neural Networks for NLP**

Proceedings of the Workshop

October 29, 2026

The BlackboxNLP organizers gratefully acknowledge the support from the following sponsors.

NDIF



©2026 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-476-7

Message from the Organizing Committee

As researchers achieve unprecedented technological breakthroughs in natural language processing, the need to understand the systems underlying these advances is more pertinent than ever. BlackboxNLP, now in its ninth iteration, has played an important role in bringing together scholars from a diverse range of backgrounds in order to rigorously study the behavior, representations, and computations of “black-box” neural network models. Our workshop showcases original, cutting-edge research on topics including but not limited to:

- Adapting and applying analysis techniques from other disciplines, such as neuroscience, to analyze high-dimensional vector representations in artificial neural networks.
- Examining model performance on simplified or formal languages.
- Proposing architectural modifications to increase models’ interpretability.
- Testing if interpretable information can be decoded from internal representations.
- Open-source tools for analysis, visualization, or explanation to democratize access to interpretability techniques in NLP.
- Meta-evaluation of analysis methods to assess their validity.
- Understanding how and when language models rely on context information.
- Analysing the linguistic properties captured by contextualised word representations.
- Scaling up analysis methods for large language models (LLMs).
- Mechanistic interpretability, reverse engineering approaches to understanding particular properties of neural models.
- Evaluation of techniques for steering LLM output behavior.
- Uncovering the reasoning processes of LLMs.
- Understanding under the hood of memorization in LLMs.
- Insights into LLM Failures.
- Translating interpretability insights into practical solutions to address key challenges in NLP.
- Opinion pieces about the state of interpretability in NLP.

The ninth BlackboxNLP workshop will be held in Budapest, Hungary on October 29th, 2026, hosted by the Conference on Empirical Methods in Natural Language Processing (EMNLP). 45 full papers, 12 non-archival extended abstracts, 10 ARR commitments, and 19 reproducibility special track papers were accepted for in-person presentations, from a total of 312 submissions. This year’s workshop will also feature three invited talks. BlackboxNLP 2026 would not have been possible without the high-quality peer reviews submitted by our program committee, as well as the logistical assistance provided by the EMNLP organizing committee. We gratefully acknowledge financial support from our sponsors, NDIF. Our invited speakers, authors, and presenters have allowed us to put together an outstanding program for all participants to enjoy. Welcome to BlackboxNLP! We look forward to seeing you in Budapest.

Organizing Committee

Organizing Committee

Dana Arad, Technion
Yonatan Belinkov, Technion
Hanjie Chen, Rice University
Nils Feldhus, University of Groningen
Aaron Mueller, Boston University
Hosein Mohebbi, Whispp & University of Amsterdam
Gabriele Sarti, Northeastern University
Martin Tutek, University of Zagreb

Program Committee

Reviewers

Romina Aalishah, David Demitri Africa, Ajay Agarwal, Risako Ando, Fahim Anzum, Stefan Arnold, David Arps, Leila Arras, Hadi Asghari, Darpan Aswal, Wilker Aziz

Tanja Baeumel, Selene Baez Santamaria, Lorenzo Basile, Ali Basirat, Adam Belfki, Yannis Belkhter, Sunit Bhattacharya, Jelke Bloem, Sasha Boguraev, Vamshi Krishna Bonagiri

Sara Candussio, Steffen Castle, Yik Siu Chan, Anwoy Chatterjee, Eric Chen, Hakaze Cho, Lina Conti, Seana Coulson, Mark Crovella, Marko Culjak

Tejas Dahiya, Can Demircan, Chunyuan Deng, Alexey Dontsov, Varun Dutt, Nicolas Audinet de Pieuchon

Veeraraju Elluru

Zhouxiang Fang, Zhaoxin Feng, Pedro Lobato Ferreira, Sheridan Feucht, Ahmed Frikha

Tatiana Gaintseva, Diana Galvan-Sosa, Ward Gauderis, Sebastian Gerstner, Gizem Gezici, Elena Golimblevskaia, Michal Golovanevsky, Michael Eric Goodale, Ziwei Gu, Adrien Guille, Aviral Gupta, Yoav Gur-Arieh, Daniil Gurgurov

Janiça Hackenbuchner, Michael Hanna, Yixiong Hao, Uri Hasson, Arman Hatami, Benjamin Heizerling, Elize Herrewijnen, Tianze Hua, Celine Hudelot

Elsayed Issa

Connor Thomas Jerzak, Shuai Jiang, Richard Johansson, Abhinav Joshi, Fanny Jourdan

David Kaczér, Patrick Kahardipraja, Laura Kallmeyer, To Eun Kim, Marianne De Heer Kloots, Jaroslav Kopcan, Evgenii Kortukov, Anton Korznikov, Jenny Kunz

Vivek Harsha Vardhan Lakkamaneni, Carsten Lanquillon, Sewoong Lee, Victoria R Li, Jongwon Lim, Danni Liu, Yang Janet Liu, Haoyan Luo

Haotian Ma, Jianfei Ma, Iuri Macocco, Syed Naveed Mahmood, Arnau Marin-Llobet, Sara Vera Marjanovic, Harry Mayne, Andrianos Michail, Mohammad Reza Modarres, Omar Momen, Philipp Mondorf, Benedetta Muscato, Ricardo Muñoz Sánchez

Vivi Nastase, Vera Neplenbroek

Francesco Ortu, Simon Ostermann, Kentaro Ozeki

Thomas Palmeira Ferraz, Sana Pandey, Álvaro Parafita, Angelina Parfenova, Gonçalo Paulo, Van-Thuy Phi, Antonin Poché, Lukas Galke Poeh, Mattia Proietti, Samuele Punzo, Bruno Puri

Philip Quirke

Vatsal Raina, Neel Rajani, Nathanaël Carraz Rakotonirina, Michaela Regneri, Elisabetta Rocchetti, Sophie Rosset, William Rudman, Maximo Rulli, Elena Sofia Ruzzetti

Yusuke Sakai, Suchir Salhan, Daniel Scalena, Sigurd Schacht, Vera Schmitt, Stefan F. Schouten, Zhengyang Shan, Serge Sharoff, Fei Shen, Rajvee Sheth, Shikhar Shiromani, Michal Spiegel, Bo Su, Deshan Sumanathilaka, Yu Suzuki

Eric Todd

Shiva Upadhye

Michiel T. van der Meer

Andreas Waldis, Sheng-Yu Wang, Miriam Wanner, Taro Watanabe, Daniel Wurgaft

Haotian Xia

Jing Yang, Minglai Yang, Zejia You

Kerem Zaman, Chengjun Zhang

Keynote Talk

Ivan Titov

University of Edinburgh & University of Amsterdam

Keynote Talk

Sheridan Feucht

Northeastern University

Keynote Talk

Michael Hahn

Saarland University

Table of Contents

<i>import pandas as np: the numpy strikes back</i>	
Haotian Yang	1
<i>On The Effectiveness-Fluency Trade-Off In LLM Conditioning: A Systematic Study</i>	
Iuri Macocco, Pau Rodriguez, Arno Blaas, Luca Zappella, Marco Baroni and Xavier Suau	15
<i>Cultural Binding Heads in Language Models</i>	
Avrile Floro and Luca Benedetto	34
<i>Lost in Specialization: Embedding Geometry and Retrieval Quality Across Specialized Corpora</i>	
Cornelia Ferner and Stefan Wegenkittl	49
<i>Compact Language, Complex Model Shifts: How and Where Ambiguity and Underspecification Affect LLMs</i>	
Michaela Regneri, Nina Scheller and Sören Laue	58
<i>How Much Do Modern Embeddings Reveal? Inversion Attacks on Commercial Embedding Models</i>	
Kareem Elaasar and Cristina Lopes	76
<i>Does the Greater-Than Circuit Generalize? A Reproducibility Study Across Models and Numerical Entailment</i>	
Vinay Setty	90
<i>A circuit for predicting hierarchical structure in-context in Large Language Models</i>	
Tankred Saanum, Can Demircan, Samuel J. Gershman and Eric Schulz	103
<i>Inside self-reflection: the geometry in multi-dimensional scoring of metacognitive structure</i>	
Praveen Suthaharan, Philip Robert Corlett and Yuen-Siang Ang	122
<i>Data Attribution of Emergent Misalignment with Persona Features</i>	
Clemens Vetter, David Kaczér, Lucie Flek and Florian Mai	136
<i>Specialized Attention Heads in Early Word Reconstruction</i>	
Gustaf Gren, Robert Östling and Murathan Kurfali	166
<i>From Plausible to Actionable: A Position on LLM Self-Explanations</i>	
Elize Herrewijnen, Benedetta Muscato, Gizem Gezici and Fosca Giannotti	178
<i>How Far Do SAE Features Persist Across LLM Layers? An Empirical Study with an FFN Pruning Application</i>	
Vaishakh Sreekanth Menon and Asif Ekbali	189
<i>Decodability Is Not Computation: Examining Probing Claims</i>	
Kabir Singh Grewal	201
<i>Source Marking in Language Models: A Controlled Framework for Acquisition, Cue Sensitivity, and Reasoning</i>	
Daniil Huryn	220
<i>Lost in Perception: Isolating Perceptual and Reasoning Failures in Multimodal Physics and Geometry Reasoning</i>	
Raj Jaiswal, Sree Krishna Uppalapati, Dhruvkumar Patel, Ria Khatoniar, Tanuja Ganu and Rajiv Ratn Shah	232

<i>Probing for Knowledge Attribution in Large Language Models</i>	
Ivo Brink, Alexander Boer and Dennis Ulmer	269
<i>Induction in Both Directions: A Mechanistic Analysis of In-Context Learning in Masked Diffusion Language Models</i>	
Andy Eduard Catruna and Emilian Radoi	294
<i>What Does BERT Know about Linguistic Features: Probing Finnish Model for Text Difficulty</i>	
Yiheng Wu, Jue Hou and Roman Yangarber	304
<i>Where Animacy Lives in Large Language Models: Tracing the Circuits of the Animacy Concept</i>	
Samuele Punzo, Giovanni Cinà and Sandro Pezzelle	316
<i>Do Chess Explanations Reflect Model Decisions? Behavioral and Token-Level Tests of LLM Reasoning Faithfulness</i>	
Angelina Parfenova	349
<i>Do Large Language Models appeal to the Senses? Narrative Embodiment in Stories told by Llama and Qwen</i>	
Stefan Arnold and Rene Gröbner	368
<i>Reverse Engineering the Reversal Curse: What Entity Order in Pretraining Does and Does Not Explain</i>	
Ivan Samarskyi and Ondřej Bojar	374
<i>Can Interpretation Predict Behavior on Unseen Data?</i>	
Victoria R Li, Jenny Kaufmann, Tian Qin, Martin Wattenberg, David Alvarez-Melis and Naomi Saphra	384
<i>Multiple Recovery Paths in a Lesioned Transformer: Internal Rerouting and Readout Remapping</i>	
Lorenzo Cannata and Uri Hasson	408
<i>Testing the Limits of Truth Directions in LLMs</i>	
Angelos Poulis, Mark Crovella and Evimaria Terzi	419
<i>Forcing a Letter Suppresses Refusal: Mechanistic Evidence from WMDP</i>	
Anna Antipova, Stepan Epifantsev, Oleg Chernikov, Leon Strelkov and Alexey Dontsov	448
<i>Rewired or Gated? How Instruction Tuning Shapes Knowledge-Conflict Circuits in LLMs</i>	
Shubham Santosh Pandere, Gautam Ranka, Ritika Varshney, Navya Deshmukh, Roushni Sareen and Roshan Kumar Singh	465
<i>Belief or Circuitry? Causal Evidence for In-Context Graph Learning</i>	
Katharine Kowalyszyn, Timothy Duggan, Daniel Little and Michael C Hughes	477
<i>Does Out-of-Sight Equal Out-of-Mind in CoT Monitorability?</i>	
Pedro Lobato Ferreira, Wilker Aziz and Ivan Titov	487
<i>Linear Layer Repair: Residual-Controlled Layer Profiling for Fault-Tolerant Transformer Inference</i>	
Jiya Sheetal Rathi and André Bauer	514
<i>Pitfalls in Evaluating Interpretability Agents</i>	
Tal Haklay, Nikhil Prakash, Sana Pandey, Antonio Torralba, Aaron Mueller, Jacob Andreas, Tamar Rott Shaham and Yonatan Belinkov	523
<i>Zipf Made Me Do It: A Causal Account of Outlier Dimensions</i>	
Kartik Sharma, Shreshth Rai, Nikunj Gupta and Dinesh K Vishwakarma	561

<i>Are Arithmetic Heuristic Neurons Form-Invariant? A Mechanistic Analysis of Symbols, Text, and Code in LLMs</i>	Sharath Naganna, Tanvir Ahmed Sijan and Uddipta Kalita	573
<i>A Multi-Dimensional Evaluation of Explainability in Media Bias Detection</i>	Ting Chen, Chengjun Zhang, Benjamin M. Ampel and Sagar Samtani	590
<i>Automatically Interpreting Attribution Graphs via Probe Prompting</i>	Giuseppe Birardi and Gonalo Paulo	603
<i>Sampling Reveals Style: Unsupervised, Training-Free Discovery of Prompt-Conditional Stylistic Axes in LLM Activations</i>	Ajit Mallavarapu and Ziwei Gu	639
<i>Compositional Multilingual and Behavioral Attribute Steering</i>	Hyun Gu Kang, Daniil Gurgurov, Tanja Baeumel, Josef Van Genabith and Simon Ostermann	655
<i>Are XAI Metrics Meaningful Proxies for Human Judgment? Evaluation Explanation Quality in Claim Matching</i>	Jaroslav Kopcan, uboř Kriř, Qiwei Peng, Miroslav Blstak and Vera Schmitt	677
<i>Your Judge Inherits Your Prompt: In-Context Label Bias in an LLM-as-Judge Is Directional, Not Symmetric</i>	Abdullah Ahmad, Ahmed J O Ahmed and Luca Citi	696
<i>Attend or Perish: Benchmarking Attention on Algorithmic Reasoning</i>	Michal Spiegel, Michal řtefnik, Marek Kadlık, Valeria Ruscio, Martin Fajık and Josef Kuchař	707
<i>Dissecting Lie Expectancy in Language Models</i>	Xingyuan Zhao and Seana Coulson	726
<i>The Geometry of Political Abstraction in Large Language Models</i>	No Durandard, Saurabh Dhawan and Thierry Poibeau	739
<i>How Much Sonority Structure Do Phonotactic Language Models Learn from Distribution?</i>	Jia Can Xu and Mario Giulianelli	761
<i>Do ICL and Finetuned Misalignment Share a Causal Component? Evidence from Risky Financial Advice</i>	Albrave Raphael Albayda	783
<i>From Rage to Refusal : Using Social-Media Data to Steer Hostility-Related Affect in LLMs</i>	Gayatri Krishnakumar, Maximilian Puelma Touzel, Zachary Yang, Jean-Franois Godbout and Reihaneh Rabbany	795
<i>Where Does Authorship Signal Emerge in Encoder-Based Language Models?</i>	Francis Kulumba, Guillaume Vimont, Laurent Romary and Florian Cafiero	818
<i>A Survey on the Linear Representation Hypothesis</i>	Sewoong Lee, Marc E. Canby, Ikhyun Cho and Julia Hockenmaier	835
<i>The Latin Substrate: How Language Models Represent and Mediate Script Choice</i>	Daniil Gurgurov, Alan Saji, Katharina Trinley, Josef Van Genabith and Simon Ostermann . . .	852
<i>MACE-XAI: A Multi-Agent Crowd Evaluation Framework for Robust Natural Language Explanations</i>	Renato Okabayashi Miyaji	879

<i>Automated Taxonomy Discovery for Interpretable Diagnosis of Multi-Hop Question Answering Datasets</i>	
Ryoji Osaki, Masaki Asada and Makoto Miwa	892
<i>Explaining the Influence of Nonverbal Streams Across Sentence Types</i>	
Lipisha Chaudhary and Ifeoma Nwogu	919
<i>MERGE: Minimal Expression-Replacement GEneralization Test for Natural Language Inference</i>	
Mădălina Zgreabă, Tejaswini Deoskar and Lasha Abzianidze	937
<i>The Geometry of Metaphor: Probing and Steering Metaphor Representations in Large Language Models</i>	
Shihui Li, Mette Friis Andersen and Martha Lewis	965
<i>When Context Misleads: Surprisal, Energy and Attention Entropy as Metrics of Coherence Illusions in LLMs</i>	
Ece Takmaz, Nitin Kumar, Li Kloostera and Jakub Dotlacil	986
<i>Present but Barely Used: Reproducing the Emergent Othello World Model in a Masked Diffusion Language Model</i>	
Éric Jacopin	1007
<i>Do Function Vectors and Task Vectors Generalize? A Reproducibility Study Across Model Families and Scale</i>	
Serendeeep Rudraraju	1017
<i>Memory vs. Context? Influential Factors of Factual Recall in Language Models</i>	
Guilhem Fouilhé, Nicholas Asher and Philippe Muller	1029
<i>Demystifying Entropy-based Selection for Chain-of-Thought Compression in Large Reasoning Models</i>	
Sara Candussio, Daniel Scalena, Elisabetta Fersini, Luca Bortolussi, Malvina Nissim and Gabriele Sarti	1047
<i>Limitations of Automated Simulatability: LLM Simulators Can Bypass Explanations</i>	
Antonin Poché, Fanny Jourdan, Nils Feldhus, Qianli Wang, Jing Yang, Simon Ostermann, Nicholas Asher, Philippe Muller and Vera Schmitt	1067
<i>Recurrence Is Not Enough: Causally Validating Multilingual SAE Translation Features in Gemma 2 and 3</i>	
Giang Son Nguyen, Nhi Ngoc-Yen Nguyen, Wray Buntine and Dung D. Le	1100
<i>Reproducing and Evaluating the Generalizability of Subliminal Learning in Open-Weight Models</i>	
Daan van der Weijden, Nathan Brack and Selene Baez Santamaria	1113
<i>Can Sparse AutoEncoder-Based Interpretability Demonstrate Geographic Knowledge Beyond Orthographic Features?</i>	
Thomas Hitchon, Rémy Decoupes and Adrien Guille	1121
<i>Inoculate or Reflect? Two Training Interventions Under Prompting, Steering, and Patching</i>	
Ayesha Imran and Muhammad Aaliyan	1128
<i>Is Refusal Direction Universal across Low-Resourced Languages?</i>	
Anastasia Ivanova	1143
<i>Scale-Invariant Normalization Hides Activation Scaling: Why Patch Site Matters for Demographic Bias Localization in Clinical Text Generation</i>	
Yoana Tsoneva, Fedor Splitt, Awsam Agbaria, Alexandra Tsiakalou, Jing Yang and Vera Schmitt	1155

<i>Probing Without Labels: Contrast-only Reproductions of Interpretability Results</i>	
Stefan F. Schouten and Peter Bloem	1167
<i>Reproducing the Results, But not the Interpretations: Evidence for Artifacts in Automatic Creativity Scoring with LLMs</i>	
Carl Hermany, Omar Momen and Sina Zarrieß	1184
<i>Claude reinvestigates herself: A replication of different types of syntactic agreement recruit the same units in LLMs</i>	
Elena Polyakova, Shivan Arora, Suchir Salhan and Paula Buttery	1196
<i>Are Latent Reasoning Tokens Unnecessary for Model Performance?</i>	
Austin Long, Colin Paterson and Jing Xu	1223
<i>Same Sink, Different Plumbing: Reproducing and Refining the Mechanism of Attention Sinks</i>	
Abdullah Al Zubayer, Syed Naveed Mahmood, Md Rezaur Rahman Bhuiyan, Mir Md Mohiuddin Abrar, Jareen Tasneem Khondaker, Tasfia Zaman and Farig Sadeque	1233
<i>Development of Linear Truth Encodings in Language Models: A Replication Study</i>	
Mingchuan Zou, Suchir Salhan and Paula Buttery	1248
<i>Stress-Testing Random-Vector Controls for Activation Steering</i>	
Behnood Mashayekhi, Parsa Hosseini and Arian Amiramjadi	1264
<i>Reproducing Transparent and Scrutable Recommendations: Exploring Open-Weight Models via Natural-Language User Profiles</i>	
Noah Mamié and Laurin van den Bergh	1273

Program

Thursday, October 29, 2026

09:00 - 09:10	<i>Opening Remarks</i>
09:10 - 09:50	<i>Keynote 1</i>
09:50 - 10:30	<i>Keynote 2</i>
10:30 - 11:00	<i>Coffee break</i>
11:00 - 12:00	<i>Oral Session 1</i>
12:00 - 13:00	<i>Poster Session 1</i>
13:00 - 14:00	<i>Lunch</i>
14:00 - 14:30	<i>Oral Session 2</i>
14:30 - 15:15	<i>Keynote 3</i>
15:15 - 15:30	<i>Best Paper</i>
15:30 - 15:40	<i>Closing Remarks and Awards</i>
15:40 - 16:00	<i>Coffee break</i>
16:00 - 17:00	<i>Poster Session 2</i>

import pandas as np: the numpy strikes back
Haotian Yang

On The Effectiveness-Fluency Trade-Off In LLM Conditioning: A Systematic Study
Iuri Macocco, Pau Rodriguez, Arno Blaas, Luca Zappella, Marco Baroni and Xavier Suau

Thursday, October 29, 2026 (continued)

Cultural Binding Heads in Language Models

Avrile Floro and Luca Benedetto

Lost in Specialization: Embedding Geometry and Retrieval Quality Across Specialized Corpora

Cornelia Ferner and Stefan Wegenkittl

Compact Language, Complex Model Shifts: How and Where Ambiguity and Underspecification Affect LLMs

Michaela Regneri, Nina Scheller and Sören Laue

How Much Do Modern Embeddings Reveal? Inversion Attacks on Commercial Embedding Models

Kareem Elaasar and Cristina Lopes

Does the Greater-Than Circuit Generalize? A Reproducibility Study Across Models and Numerical Entailment

Vinay Setty

A circuit for predicting hierarchical structure in-context in Large Language Models

Tankred Saanum, Can Demircan, Samuel J. Gershman and Eric Schulz

Inside self-reflection: the geometry in multi-dimensional scoring of metacognitive structure

Praveen Suthaharan, Philip Robert Corlett and Yuen-Siang Ang

Data Attribution of Emergent Misalignment with Persona Features

Clemens Vetter, David Kaczér, Lucie Flek and Florian Mai

Specialized Attention Heads in Early Word Reconstruction

Gustaf Gren, Robert Östling and Murathan Kurfali

From Plausible to Actionable: A Position on LLM Self-Explanations

Elize Herrewijnen, Benedetta Muscato, Gizem Gezici and Fosca Giannotti

How Far Do SAE Features Persist Across LLM Layers? An Empirical Study with an FFN Pruning Application

Vaishakh Sreekanth Menon and Asif Ekbal

Thursday, October 29, 2026 (continued)

Decodability Is Not Computation: Examining Probing Claims

Kabir Singh Grewal

Source Marking in Language Models: A Controlled Framework for Acquisition, Cue Sensitivity, and Reasoning

Daniil Huryn

Lost in Perception: Isolating Perceptual and Reasoning Failures in Multimodal Physics and Geometry Reasoning

Raj Jaiswal, Sree Krishna Uppalapati, Dhruvkumar Patel, Ria Khatoniar, Tanuja Ganu and Rajiv Ratn Shah

Probing for Knowledge Attribution in Large Language Models

Ivo Brink, Alexander Boer and Dennis Ulmer

Induction in Both Directions: A Mechanistic Analysis of In-Context Learning in Masked Diffusion Language Models

Andy Eduard Catruna and Emilian Radoi

What Does BERT Know about Linguistic Features: Probing Finnish Model for Text Difficulty

Yiheng Wu, Jue Hou and Roman Yangarber

Where Animacy Lives in Large Language Models: Tracing the Circuits of the Animacy Concept

Samuele Punzo, Giovanni Cinà and Sandro Pezzelle

Do Chess Explanations Reflect Model Decisions? Behavioral and Token-Level Tests of LLM Reasoning Faithfulness

Angelina Parfenova

Do Large Language Models appeal to the Senses? Narrative Embodiment in Stories told by Llama and Qwen

Stefan Arnold and Rene Gröbner

Reverse Engineering the Reversal Curse: What Entity Order in Pretraining Does and Does Not Explain

Ivan Samarskyi and Ondřej Bojar

Can Interpretation Predict Behavior on Unseen Data?

Victoria R Li, Jenny Kaufmann, Tian Qin, Martin Wattenberg, David Alvarez-Melis and Naomi Saphra

Thursday, October 29, 2026 (continued)

Multiple Recovery Paths in a Lesioned Transformer: Internal Rerouting and Readout Remapping

Lorenzo Cannata and Uri Hasson

Testing the Limits of Truth Directions in LLMs

Angelos Poulis, Mark Crovella and Evimaria Terzi

Forcing a Letter Suppresses Refusal: Mechanistic Evidence from WMDP

Anna Antipova, Stepan Epifantsev, Oleg Chernikov, Leon Strelkov and Alexey Dontsov

Rewired or Gated? How Instruction Tuning Shapes Knowledge-Conflict Circuits in LLMs

Shubham Santosh Pandere, Gautam Ranka, Ritika Varshney, Navya Deshmukh, Roushni Sareen and Roshan Kumar Singh

Belief or Circuitry? Causal Evidence for In-Context Graph Learning

Katharine Kowalyshyn, Timothy Duggan, Daniel Little and Michael C Hughes

Does Out-of-Sight Equal Out-of-Mind in CoT Monitorability?

Pedro Lobato Ferreira, Wilker Aziz and Ivan Titov

Linear Layer Repair: Residual-Controlled Layer Profiling for Fault-Tolerant Transformer Inference

Jiya Sheetal Rathi and André Bauer

Pitfalls in Evaluating Interpretability Agents

Tal Haklay, Nikhil Prakash, Sana Pandey, Antonio Torralba, Aaron Mueller, Jacob Andreas, Tamar Rott Shaham and Yonatan Belinkov

Zipf Made Me Do It: A Causal Account of Outlier Dimensions

Kartik Sharma, Shreshth Rai, Nikunj Gupta and Dinesh K Vishwakarma

Are Arithmetic Heuristic Neurons Form-Invariant? A Mechanistic Analysis of Symbols, Text, and Code in LLMs

Sharath Naganna, Tanvir Ahmed Sijan and Uddipta Kalita

A Multi-Dimensional Evaluation of Explainability in Media Bias Detection

Ting Chen, Chengjun Zhang, Benjamin M. Ampel and Sagar Samtani

Thursday, October 29, 2026 (continued)

Automatically Interpreting Attribution Graphs via Probe Prompting

Giuseppe Birardi and Gonalo Paulo

Sampling Reveals Style: Unsupervised, Training-Free Discovery of Prompt-Conditional Stylistic Axes in LLM Activations

Ajit Mallavarapu and Ziwei Gu

Compositional Multilingual and Behavioral Attribute Steering

Hyun Gu Kang, Daniil Gurgurov, Tanja Baeumel, Josef Van Genabith and Simon Ostermann

Are XAI Metrics Meaningful Proxies for Human Judgment? Evaluation Explanation Quality in Claim Matching

Jaroslav Kopcan, uboř Kriř, Qiwei Peng, Miroslav Blstak and Vera Schmitt

Your Judge Inherits Your Prompt: In-Context Label Bias in an LLM-as-Judge Is Directional, Not Symmetric

Abdullah Ahmad, Ahmed J O Ahmed and Luca Citi

Attend or Perish: Benchmarking Attention on Algorithmic Reasoning

Michal Spiegel, Michal řtefnik, Marek Kadlk, Valeria Ruscio, Martin Fajk and Josef Kuchař

Dissecting Lie Expectancy in Language Models

Xingyuan Zhao and Seana Coulson

The Geometry of Political Abstraction in Large Language Models

No Durandard, Saurabh Dhawan and Thierry Poibeau

How Much Sonority Structure Do Phonotactic Language Models Learn from Distribution?

Jia Can Xu and Mario Giulianelli

Do ICL and Finetuned Misalignment Share a Causal Component? Evidence from Risky Financial Advice

Albrave Raphael Albayda

From Rage to Refusal : Using Social-Media Data to Steer Hostility-Related Affect in LLMs

Gayatri Krishnakumar, Maximilian Puelma Touzel, Zachary Yang, Jean-Franois Godbout and Reihaneh Rabbany

Thursday, October 29, 2026 (continued)

Where Does Authorship Signal Emerge in Encoder-Based Language Models?

Francis Kulumba, Guillaume Vimont, Laurent Romary and Florian Cafiero

A Survey on the Linear Representation Hypothesis

Sewoong Lee, Marc E. Canby, Ikhyun Cho and Julia Hockenmaier

The Latin Substrate: How Language Models Represent and Mediate Script Choice

Daniil Gurgurov, Alan Saji, Katharina Trinley, Josef Van Genabith and Simon Ostermann

MACE-XAI: A Multi-Agent Crowd Evaluation Framework for Robust Natural Language Explanations

Renato Okabayashi Miyaji

Automated Taxonomy Discovery for Interpretable Diagnosis of Multi-Hop Question Answering Datasets

Ryoji Osaki, Masaki Asada and Makoto Miwa

Explaining the Influence of Nonverbal Streams Across Sentence Types

Lipisha Chaudhary and Ifeoma Nwogu

MERGE: Minimal Expression-Replacement Generalization Test for Natural Language Inference

Mădălina Zgreabă, Tejaswini Deoskar and Lasha Abzianidze

The Geometry of Metaphor: Probing and Steering Metaphor Representations in Large Language Models

Shihui Li, Mette Friis Andersen and Martha Lewis

When Context Misleads: Surprisal, Energy and Attention Entropy as Metrics of Coherence Illusions in LLMs

Ece Takmaz, Nitin Kumar, Li Kloostera and Jakub Dotlacil

Present but Barely Used: Reproducing the Emergent Othello World Model in a Masked Diffusion Language Model

Éric Jacopin

Do Function Vectors and Task Vectors Generalize? A Reproducibility Study Across Model Families and Scale

Serendeeep Rudraraju

Thursday, October 29, 2026 (continued)

Memory vs. Context? Influential Factors of Factual Recall in Language Models

Guilhem Fouilhé, Nicholas Asher and Philippe Muller

Demystifying Entropy-based Selection for Chain-of-Thought Compression in Large Reasoning Models

Sara Candussio, Daniel Scalena, Elisabetta Fersini, Luca Bortolussi, Malvina Nissim and Gabriele Sarti

Limitations of Automated Simulatability: LLM Simulators Can Bypass Explanations

Antonin Poché, Fanny Jourdan, Nils Feldhus, Qianli Wang, Jing Yang, Simon Ostermann, Nicholas Asher, Philippe Muller and Vera Schmitt

Recurrence Is Not Enough: Causally Validating Multilingual SAE Translation Features in Gemma 2 and 3

Giang Son Nguyen, Nhi Ngoc-Yen Nguyen, Wray Buntine and Dung D. Le

Reproducing and Evaluating the Generalizability of Subliminal Learning in Open-Weight Models

Daan van der Weijden, Nathan Brack and Selene Baez Santamaria

Can Sparse AutoEncoder-Based Interpretability Demonstrate Geographic Knowledge Beyond Orthographic Features?

Thomas Hitchon, Rémy Decoupes and Adrien Guille

Inoculate or Reflect? Two Training Interventions Under Prompting, Steering, and Patching

Ayesha Imran and Muhammad Aaliyan

Is Refusal Direction Universal across Low-Resourced Languages?

Anastasia Ivanova

Scale-Invariant Normalization Hides Activation Scaling: Why Patch Site Matters for Demographic Bias Localization in Clinical Text Generation

Yoana Tsoneva, Fedor Splitt, Awsam Agbaria, Alexandra Tsiakalou, Jing Yang and Vera Schmitt

Probing Without Labels: Contrast-only Reproductions of Interpretability Results

Stefan F. Schouten and Peter Bloem

Reproducing the Results, But not the Interpretations: Evidence for Artifacts in Automatic Creativity Scoring with LLMs

Carl Hermany, Omar Momen and Sina Zarrieß

Thursday, October 29, 2026 (continued)

Claude reinvestigates herself: A replication of different types of syntactic agreement recruit the same units in LLMs

Elena Polyakova, Shivan Arora, Suchir Salhan and Paula Buttery

Are Latent Reasoning Tokens Unnecessary for Model Performance?

Austin Long, Colin Paterson and Jing Xu

Same Sink, Different Plumbing: Reproducing and Refining the Mechanism of Attention Sinks

Abdullah Al Zubayer, Syed Naveed Mahmood, Md Rezaur Rahman Bhuiyan, Mir Md Mohiuddin Abrar, Jareen Tasneem Khondaker, Tasfia Zaman and Farig Sadeque

Development of Linear Truth Encodings in Language Models: A Replication Study

Mingchuan Zou, Suchir Salhan and Paula Buttery

Stress-Testing Random-Vector Controls for Activation Steering

Behnood Mashayekhi, Parsa Hosseini and Arian Amiramjadi

Reproducing Transparent and Scrutable Recommendations: Exploring Open-Weight Models via Natural-Language User Profiles

Noah Mamié and Laurin van den Bergh

16:00 - 17:00	<i>Helical Representations of Turn Structure in an LLM</i>
16:00 - 17:00	<i>Explaining GAND: A Resource on Gender-Ambiguous Natural Data & Contrastive Attribution</i>
16:00 - 17:00	<i>Geometric Representations of African Languages: A Regional Semantic Hub and Cultural Steering</i>
16:00 - 17:00	<i>Do Hyper-Connections Actually Use Multiple Residual Streams? Diagnosing and Mitigating Stream Collapse</i>
16:00 - 17:00	<i>Simulating Language Impairment via Causal Head Ablation</i>
16:00 - 17:00	<i>The Sum Is Less Than Its Parts: Decomposing the Attention Write by Source Detects Contextual Hallucinations</i>
16:00 - 17:00	<i>Temporal Preference Concepts and Their Functions in a Large Language Model</i>

Thursday, October 29, 2026 (continued)

- 16:00 - 17:00 *Weird Generalization is Weirdly Brittle*
- 16:00 - 17:00 *Do LLM Explanations Align with the Brain? Evidence from Attribution and Layer Conductance*
- 16:00 - 17:00 *Processing Cross-serial Dependencies with Decoder Large Language Models*
- 16:00 - 17:00 *SHIFT Does Not Generalize Reliably*
- 16:00 - 17:00 *Belief-Reality Separation Lives in Routing over a Shared Value Slot in Language Models*