Organizing Committee

General Chair

Yi Yang, ASAPP
Aida Davani, Google Research
Avi Sil, IBM
Anoop Kumar, Capital One
Program Committee

Reviewers

Mohamed Abdelhady, Amazon
Sachin Agarwal, Apple
Prabhat Agarwal, Pinterest, Inc.
Alan Akbik, Humboldt Universität Berlin
Burak Aksar
Mohamed AlTantawy, Agolo
Enrique Henestroza Anguiano
Ankit Arun
AiTi Aw, I2R
Kfir Bar, College of Management
Leslie Barrett, Bloomberg, LP
Emre Barut, Amazon
Daniel Bauer, Columbia University
Frederic Bechet, Académie d’Aix-Marseille
Kasturi Bhattacharjee, Pryon and AWS AI
Trung Bui, Adobe Research
Sai Kiran Burle
Aoife Cahill, Dataminr
Sarah C Campbell, Amazon Alexa
Thiago Castro Ferreira, Universidade Federal de Minas Gerais
Sourish Chaudhuri
John Chen, Department of Speech and Natural Language Research, Interactions LLC
Luoxin Chen, Amazon
Jiangning Chen, UKG
Pengxiang Cheng, Bloomberg
Justin Chiu, Rakuten Institute of Technology, The University of Tokyo
Jaegul Choo, Korea Advanced Institute of Science and Technology
Deborah A. Dahl, Open Voice Interoperability Initiative and Conversational Technologies
Marina Danilevsky, International Business Machines
Aswarth Abhilash Dara
Anirban Das, Capital One
Vivek Datla, Capital One
Rahul Divekar, Educational Testing Service
Shuyan Dong, Facebook
Li Dong, Amazon
Matthew T. Dunn
Matthias Eck, Carnegie Mellon University
Lilach Eden
Wassim El-Hajj, American University of Beirut
Aparna Elangovan, Amazon
David Elson, Google
Ramy Eskander, Google
Michael Flor, Educational Testing Service
Lisheng Fu, Comcast
Aram Galstyan, Information Sciences Institute, University of Southern California and Amazon Alexa
Radhika Gaonkar
Jose Garrido Ramas
Diman Ghazi
Anmol Goel, Technische Universität Darmstadt
Olga Golovneva, Facebook
Tong Guo
Ankush Gupta, IBM India Research Lab
Dilek Hakkani-Tur, University of Illinois at Urbana-Champaign
Benjamin Han, Apple
Hua He
Sanjika Hewawitharana, eBay Inc.
Wonseok Hwang, University of Seoul and LBox Co., Ltd.
Leslie Ikemoto
Alankar Jain
Rosie Jones, Spotify
Mohammad Kachuee, Amazon
Anup K. Kalia
Hidetaka Kamigaito, Division of Information Science, Nara Institute of Science and Technology
Jun Seok Kang
Damianos Karakos
Yannis Katsis, International Business Machines
Nikhil Khani, Google
Saurabh Khanwalkar, Course Hero Inc.
Kunho Kim, Microsoft
Geewook Kim, NAVER Cloud and KAIST
Sun Kim, Naver
Rajasekar Krishnamurthy, Adobe Systems
Vinayshekhar Bannihatti Kumar, Amazon
Anjishnu Kumar
Sanjeev Kumar
Sarasi Lalithsena
Brian Lester, Department of Computer Science, University of Toronto and Google
Yulong Li, IBM, International Business Machines
Zhouhan Lin, Shanghai Jiao Tong University
Antonie Lin, Amazon
Xuye Liu
Petr Lorenc
Liang Ma, Dataminr
Fred Mailhot, Dialpad, Inc.
Lorenzo Malandrí, University of Milan - Bicocca
Yuval Marton, Genentech and University of Washington
Yuji Matsumoto, RIKEN Center for Advanced Intelligence Project
Chandresh Kumar Maurya, Indian Institute of Technology, Indore
Arne Mauser, Snowflake
David D. McDonald
Kartik Mehta, Amazon
Fabio Mercorio, University of Milan - Bicocca
Margot Mieskes, University of Applied Sciences Darmstadt
Nyalleng Moorosi, Distributed AI Research
Sidharth Mudgal, Google
Matthew Mulholland, Educational Testing Service
Deepak Muralidharan, Apple
Prasanna Muthukumar
Varun Nagaraj Rao, Princeton University
Jinseok Nam, Amazon
Nobal B. Niraula, Boeing Research & Technology
Navid Nobani
Sergio Oramas, SiriusXM / Pandora
Laurel Orr, Computer Science Department, Stanford University
Feifei Pan
Taiwoo Park, NAVER Search US
Cheoneum Park, Hyundai Motor Group
Dookun Park
Abhay Dutt Paroha
Ioannis Partalas
Sangameshwar Patil, Indian Institute of Technology, Madras and Tata Consultancy Services Limited, India
Sachin Pawar
Stephan Petitz, Apple
Xujun Peng, Amazon
Pradyot Prakash, Facebook
Radityo Eko Prasojo, Rukita
Stephen Pulman, Apple
Haode Qi
Long Qin, Alibaba Group
Elio Querze
Nitin Ramrakhiyani, International Institute of Information Technology Hyderabad and Tata Consultancy Services Limited, India
Shihao Ran
Vivek Kumar Rangarajan Sridhar
Nikhil Rasiwasia, Facebook
Ehud Reiter, University of Aberdeen
Giuseppe Riccardi, University of Trento
Alicia Sagae, Amazon
Avneesh Saluja, Netflix
Thomas Schaaf
Jonathan Schler, Holon Institute of Technology
Frank Seide
Jaydeep Sen
Shubhashis Sengupta
Igor Shalyminov, Amazon
Mingyue Shang, Amazon
Michal Shmueli-Scheuer
Lei Shu, Google
Svetlana Stoyanchev, Toshiba Research Europe
Marek Suppa, Comenius University in Bratislava
Sandesh Swamy, Amazon
Narges Tabari, Amazon
Joel R. Tetreault
Sudarshan R. Thitte, International Business Machines
Christoph Tillmann
Giuliano Tortoreto
Isabel Trancoso, Instituto Superior Técnico
Aashka Trivedi, International Business Machines
Keith Trnka
Morgan Ulinski, Soar Technology, LLC
David Uthus, Google
Vidya Venkiteswaran
Ngoc Phuoc An Vo, International Business Machines
Dakuo Wang, Northeastern University
Tong Wang, Amazon
Kyle Williams, Microsoft
Ziyun Xu
Ziyun Xu
Xiao Yang, Facebook and Facebook
Jin Yeong Yim
Keunwoo Peter Yu, University of Michigan - Ann Arbor
Qingkai Zeng, University of Notre Dame
Ke Zhang, Dataminr, inc
Yichao Zhou, Google
Xiliang Zhu, Dialpad Inc.
Chenyang Zhu
Hila Weisman Zohar
Bowie Zou, A*STAR
Table of Contents

**HPipe: Large Language Model Pipeline Parallelism for Long Context on Heterogeneous Cost-effective Devices**  
Ruilong Ma, Xiang Yang, Jingyu Wang, Qi Qi, Haifeng Sun, Jing Wang, Zirui Zhuang and Jianxin Liao ................................................................. 1

**Lossless Acceleration of Large Language Model via Adaptive N-gram Parallel Decoding**  
Jie Ou, Yueming Chen and Prof. Wenhong Tian .................................................. 10

**SOLAR 10.7B: Scaling Large Language Models with Simple yet Effective Depth Up-Scaling**  
Sanghoon Kim, Dahyun Kim, Chanjun Park, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeoung Kim, Yungi Kim, Hyoengju Lee, Jiho Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee and Sunghun Kim ................. 23

**UINav: A Practical Approach to Train On-Device Automation Agents**  
Wei Li, Fu-Lin Hsu, William E Bishop, Folawiyo Campbell-Ajula, Max Lin and Oriana Riva . 36

**Efficiently Distilling LLMs for Edge Applications**  
Achintya Kundu, Yu Chin Fabian Lim, Aaron Chew, Laura Wynter, Penny Chong and Rhui Dih Lee ................................................................. 52

**Modeling and Detecting Company Risks from News**  
Jiaxin Pei, Soumya Vadlamannati, Liang-Kang Huang, Daniel Preotiuc-Pietro and Xinyu Hua 63

**Multiple-Question Multiple-Answer Text-VQA**  
Peng Tang, Srikar Appalaraju, R. Mannathu, Yusheng Xie and Vijay Mahadevan ............... 73

**An NLP-Focused Pilot Training Agent for Safe and Efficient Aviation Communication**  
Xiaochen Liu, Bowei Zou and AiTi Aw ................................................................. 89

**Visual Grounding for User Interfaces**  
Yijun Qian, Yujie Lu, Alexander G Hauptmann and Oriana Riva .................................... 97

**Prompt Tuned Embedding Classification for Industry Sector Allocation**  
Valentin Leonhard Buchner, Lele Cao, Jan-Christoph Kalo and Vilhelm Von Ehrenheim .... 108

**REXEL: An End-to-end Model for Document-Level Relation Extraction and Entity Linking**  
Nacime Bouziani, Shubhi Tyagi, Joseph Fisher, Jens Lehmann and Andrea Pierleoni .......... 119

**Conformer-Based Speech Recognition On Extreme Edge-Computing Devices**  
Mingbin Xu, Alex Jin, Sicheng Wang, Mu Su, Tim Ng, Henry Mason, Shiyi Han, Zhihong Lei, Yaqiao Deng, Zhen Huang and Mahesh Krishnamoorthy ........................................... 131

**Generating Signed Language Instructions in Large-Scale Dialogue Systems**  
Mert Inan, Katherine Atwell, Anthony Sicilia, Lorna Quandt and Malihe Alikhani .......... 140

**Leveraging Natural Language Processing and Large Language Models for Assisting Due Diligence in the Legal Domain**  
Myeongjun Erik Jang and Gábor Stikkel ............................................................. 155

**AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators**  
Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan and Weizhu Chen .................................................. 165
An Automatic Prompt Generation System for Tabular Data Tasks
Ashlesha Akella, Abhijit Manatkar, Brijkumar Chavda and Hima Patel .............................. 191

Fighting crime with Transformers: Empirical analysis of address parsing methods in payment data
Haitham Hammami, Louis Baligand and Bojan Petrovski .................................................... 201

Language Models are Alignable Decision-Makers: Dataset and Application to the Medical Triage Domain
Brian H Hu, Bill Ray, Alice Leung, Amy Summerville, David Joy, Christopher Funk and Arslan Basharat .......................................................... 213

Reducing hallucination in structured outputs via Retrieval-Augmented Generation
Orlando Marquez Ayala and Patrice Bechard ................................................................. 228

Towards Translating Objective Product Attributes Into Customer Language
Ram Yazdi, Oren Kalinsky, Alexander Libov and Dafna Shahaf ............................................. 239

Automating the Generation of a Functional Semantic Types Ontology with Foundational Models
Sachin G Konan, Larry Rudolph and Scott Affens ............................................................... 248

Leveraging Customer Feedback for Multi-modal Insight Extraction
Sandeep Sricharan Mukku, Abinesh Kanagarajan, Pushpendu Ghosh and Chetan Aggarwal . . 266

Optimizing LLM Based Retrieval Augmented Generation Pipelines in the Financial Domain
Yiyun Zhao, Prateek Singh, Hanoz Bhatena, Bernardo Ramos, Aviral Joshi, Swaroop Gadiyaram and Saket Sharma ................................................................. 279

Scaling Up Authorship Attribution
Jacob Striebel, Abishek Edikula, Ethan Irby, Alex Rosenfeld, J. Blake Gage, Daniel Dakota and Sandra Kühler ................................................................. 295

Multimodal Contextual Dialogue Breakdown Detection for Conversational AI Models
Md Messal Monem Miah, Ulie Schnaithmann, Arushi Raghuvanshi and Youngseo Son ..... 303

Deferred NAM: Low-latency Top-K Context Injection via Deferred Context Encoding for Non-Streaming ASR
Zelin Wu, Gan Song, Christopher Li, Pat Rondon, Zhong Meng, Xavier Velez, Weiran Wang, Diamantino Caseiro, Golan Pundak, Tsendsuren Munkhdalai, Angad Chandorkar and Rohit Prabhavalkar 315

Less is More for Improving Automatic Evaluation of Factual Consistency
Tong Wang, Ninad Kulkarni and Yanjun Qi ................................................................. 324

DriftWatch: A Tool that Automatically Detects Data Drift and Extracts Representative Examples Affected by Drift
Myeongjun Erik Jang, Antonios Georgiadis, Yiyun Zhao and Fran Silavong ......................... 335

Graph Integrated Language Transformers for Next Action Prediction in Complex Phone Calls
Amin Hosseiny Marani, Ulie Schnaithmann, Youngseo Son, Akil Iyer, Manas Paldhe and Arushi Raghuvanshi ................................................................. 347

Leveraging LLMs for Dialogue Quality Measurement
Jinghan Jia, Abi Komma, Timothy Leffel, Xujun Peng, Ajay Nagesh, Tamer Soliman, Aram Galstyan and Anoop Kumar ................................................................. 359

Uncertainty Estimation in Large Language Models to Support Biodiversity Conservation
Maria Mora-Cross and Saul Calderon-Ramirez ................................................................. 368
AMA-LSTM: Pioneering Robust and Fair Financial Audio Analysis for Stock Volatility Prediction
Shengkun Wang, Taoran Ji, Jianfeng He, Mariam ALMutairi, Dan Wang, Linhan Wang, Min Zhang and Chang-Tien Lu ................................................................. 379

Tiny Titans: Can Smaller Large Language Models Punch Above Their Weight in the Real World for Meeting Summarization?
Xue-Yong Fu, Md Tahmid Rahman Laskar, Elena Khasanova, Cheng Chen and Shashi Bhushan
TN .............................................................................................................. 387

Shears: Unstructured Sparsity with Neural Low-rank Adapter Search
Juan Pablo Munoz, Jinjie Yuan and Nilesh Jain .............................................. 395

Tree-of-Question: Structured Retrieval Framework for Korean Question Answering Systems
Dongyub Lee, Younghun Jeong, Hwa-Yeon Kim, Hongyeon Yu, Seunghyun Han, Taesun Whang, Seungwoo Cho, Chanhee Lee, Gunsu Lee and Youngbum Kim ................................................................. 406

LLM-based Frameworks for API Argument Filling in Task-Oriented Conversational Systems
Jisoo Mok, Mohammad Kachuee, Shuyang Dai, Shayan Ray, Tara Taghavi and Sungroh Yoon 419

Large Language Models Encode the Practice of Medicine
Teja Kanchinadam and Gauher Shaheen ...................................................... 427

Leveraging Interesting Facts to Enhance User Engagement with Conversational Interfaces
Nikhita Vedula, Giuseppe Castellucci, Eugene Agichtein, Oleg Rokhlenko and Shervin Malmasi 437

Search Query Refinement for Japanese Named Entity Recognition in E-commerce Domain
Yuki Nakayama, Ryutaro Tatsushima, Erick Mendieta, Koji Murakami and Keiji Shinzato ... 447

EIVEN: Efficient Implicit Attribute Value Extraction using Multimodal LLM
Henry Peng Zou, Gavin Heqing Yu, Ziwei Fan, Dan Bu, Han Liu, Peng Dai, Dongmei Jia and Cornelia Caragea .............................................................................................................. 453

Exploring the Impact of Table-to-Text Methods on Augmenting LLM-based Question Answering with Domain Hybrid Data
Dehai Min, Nan Hu, Rihui Jin, Nuo Lin, Jiaoyan Chen, Yongrui Chen, Yu Li, Guilin Qi, Yun Li, Nijun Li and Qianren Wang ................................................................. 464

Solving General Natural-Language-Description Optimization Problems with Large Language Models
Jihai Zhang, Wei Wang, Siyan Guo, Li Wang, Fangquan Lin, Cheng Yang and Wotao Yin ... 483

Self-Regulated Data-Free Knowledge Amalgamation for Text Classification
Prashanth Vijayaraghavan, Hongzhi Wang, Luyao Shi, Tyler Baldwin, David Beymer and Ehsan Degan ................................................................. 491