

63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)

Short Papers

Vienna, Austria
27 July - 1 August 2025

Part 1 of 2

ISBN: 979-8-3313-2389-9

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571



Some format issues inherent in the e-media version may also appear in this print version.

Copyright© (2025) by the Association for Computational Linguistics
All rights reserved.

Printed with permission by Curran Associates, Inc. (2026)

For permission requests, please contact the Association for Computational Linguistics
at the address below.

Association for Computational Linguistics
209 N. Eighth Street
Stroudsburg, Pennsylvania 18360

Phone: 1-570-476-8006
Fax: 1-570-476-0860

acl@aclweb.org

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2633
Email: curran@proceedings.com
Web: www.proceedings.com

Table of Contents

<i>Towards LLM-powered Attentive Listener: A Pragmatic Approach through Quantity Self-Repair</i> Junlin Li, Peng Bo and Yu-Yin Hsu	1
<i>MIRAGE: Exploring How Large Language Models Perform in Complex Social Interactive Environments</i> Yin Cai, Zhouhong Gu, Zhaohan Du, Zheyu Ye, Shaosheng Cao, Yiqian xu, Hongwei Feng and Ping Chen	14
<i>Dynamic Label Name Refinement for Few-Shot Dialogue Intent Classification</i> Gyutae Park, Ingeol Baek, Byeongjeong Kim, Joongbo Shin and Hwanhee Lee	41
<i>Rethinking KenLM: Good and Bad Model Ensembles for Efficient Text Quality Filtering in Large Web Corpora</i> Yungi Kim, Hyunsoo Ha, Sukyung Lee, Jihoo Kim, Seonghoon Yang and Chanjun Park	53
<i>Automatic detection of dyslexia based on eye movements during reading in Russian</i> Anna Laurinavichyute, Anastasiya Lopukhina and David Robert Reich	59
<i>Doc-React: Multi-page Heterogeneous Document Question-answering</i> 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 and Ritwik Sinha	67
<i>ConECT Dataset: Overcoming Data Scarcity in Context-Aware E-Commerce MT</i> Mikołaj Pokrywka, Wojciech Kusa, Mieszko Rutkowski and Mikołaj Koszowski	79
<i>A Measure of the System Dependence of Automated Metrics</i> Pius Von Däniken, Jan Milan Deriu and Mark Cieliebak	87
<i>Call for Rigor in Reporting Quality of Instruction Tuning Data</i> Hyeonseok Moon, Jaehyung Seo and Heuiseok Lim	100
<i>BQA: Body Language Question Answering Dataset for Video Large Language Models</i> Shintaro Ozaki, Kazuki Hayashi, Miyu Oba, Yusuke Sakai, Hidetaka Kamigaito and Taro Watanabe	110
<i>Grounded, or a Good Guesser? A Per-Question Balanced Dataset to Separate Blind from Grounded Models for Embodied Question Answering</i> Miles Shelton, Nate Wingerd, Kritim K Rijal, Ayush Garg, Adelina Gutic, Brett Barnes and Catherine Finegan-Dollak	124
<i>Learning Sparsity for Effective and Efficient Music Performance Question Answering</i> Xingjian Diao, Tianzhen Yang, Chunhui Zhang, Weiyi Wu, Ming Cheng and Jiang Gui	136
<i>Cross-Lingual Transfer of Cultural Knowledge: An Asymmetric Phenomenon</i> Chen Zhang, Zhiyuan Liao and Yansong Feng	147
<i>Leveraging Human Production-Interpretation Asymmetries to Test LLM Cognitive Plausibility</i> Suet-Ying Lam, Qingcheng Zeng, Jingyi Wu and Rob Voigt	158
<i>Improving the Calibration of Confidence Scores in Text Generation Using the Output Distribution's Characteristics</i> Lorenzo Jaime Yu Flores, Ori Ernst and Jackie CK Cheung	172

<i>KnowShiftQA: How Robust are RAG Systems when Textbook Knowledge Shifts in K-12 Education?</i>	
Tianshi Zheng, Weihan Li, Jiaxin Bai, Weiqi Wang and Yangqiu Song	183
<i>Improving Parallel Sentence Mining for Low-Resource and Endangered Languages</i>	
Shu Okabe, Katharina Hämmerl and Alexander Fraser	196
<i>Revisiting Epistemic Markers in Confidence Estimation: Can Markers Accurately Reflect Large Language Models' Uncertainty?</i>	
Jiayu Liu, Qing Zong, Weiqi Wang and Yangqiu Song	206
<i>Limited-Resource Adapters Are Regularizers, Not Linguists</i>	
Marcell Fekete, Nathaniel Romney Robinson, Ernests Lavrinovics, Djeride Jean-Baptiste, Raj Dabre, Johannes Bjerva and Heather Lent	222
<i>LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks</i>	
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 and Alberto Testoni	238
<i>FocalPO: Enhancing Preference Optimizing by Focusing on Correct Preference Rankings</i>	
Tong Liu, Xiao Yu, Wenxuan Zhou, Jindong Gu and Volker Tresp	256
<i>Combining Domain and Alignment Vectors Provides Better Knowledge-Safety Trade-offs in LLMs</i>	
Megh Thakkar, Quentin Fournier, Matthew Riemer, Pin-Yu Chen, Amal Zouaq, Payel Das and Sarath Chandar	268
<i>Can Uniform Meaning Representation Help GPT-4 Translate from Indigenous Languages?</i>	
Shira Wein	278
<i>Subword models struggle with word learning, but surprisal hides it</i>	
Bastian Bunzeck and Sina Zarriß	286
<i>LLM as Entity Disambiguator for Biomedical Entity-Linking</i>	
Christophe Ye and Cassie S. Mitchell	301
<i>Towards Geo-Culturally Grounded LLM Generations</i>	
Piyawat Lertvittayakumjorn, David Kinney, Vinodkumar Prabhakaran, Donald Martin Jr. and Sunipa Dev	313
<i>MUSTS: MULTilingual Semantic Textual Similarity Benchmark</i>	
Tharindu Ranasinghe, Hansi Hettiarachchi, Constantin Orasan and Ruslan Mitkov	331
<i>Can Large Language Models Accurately Generate Answer Keys for Health-related Questions?</i>	
Davis Bartels, Deepak Gupta and Dina Demner-Fushman	354
<i>Literary Evidence Retrieval via Long-Context Language Models</i>	
Katherine Thai and Mohit Iyyer	369
<i>A Little Human Data Goes A Long Way</i>	
Dhananjay Ashok and Jonathan May	381
<i>Seeking Rational Demonstrations for Large Language Models: A Domain Generalization Approach to Unsupervised Cross-Domain Keyphrase Generation</i>	
Guangzhen Zhao, Yu Yao, Dechang Kong and Zhenjiang Dong	414

<i>LexKeyPlan: Planning with Keyphrases and Retrieval Augmentation for Legal Text Generation: A Case Study on European Court of Human Rights Cases</i>	
Santosh T.y.s.s and Elvin Quero Hernandez	425
<i>SynWorld: Virtual Scenario Synthesis for Agentic Action Knowledge Refinement</i>	
Runnan Fang, Xiaobin Wang, Yuan Liang, Shuofei Qiao, Jialong Wu, Zekun Xi, Ningyu Zhang, Yong Jiang, Pengjun Xie, Fei Huang and HuaJun Chen	437
<i>Enhancing Retrieval Systems with Inference-Time Logical Reasoning</i>	
Felix Faltings, Wei Wei and Yujia Bao	449
<i>Using Subtext to Enhance Generative IDRR</i>	
Zhipang Wang, Yu Hong, Weihao Sun and Guodong Zhou	464
<i>State-offset Tuning: State-based Parameter-Efficient Fine-Tuning for State Space Models</i>	
Wonjun Kang, Kevin Galim, Yuchen Zeng, Minjae Lee, Hyung Il Koo and Nam Ik Cho	474
<i>Internal and External Impacts of Natural Language Processing Papers</i>	
Yu Zhang	488
<i>An Effective Incorporating Heterogeneous Knowledge Curriculum Learning for Sequence Labeling</i>	
Xuemei Tang, Jun Wang, Qi Su, Chu-Ren Huang and Jinghang Gu	495
<i>Accelerating Dense LLMs via L0-regularized Mixture-of-Experts</i>	
Zhenyu Zhang, JiuDong Yang, Taozhaowen Taozhaowen and Meng Chen	504
<i>Do Multimodal Large Language Models Truly See What We Point At? Investigating Indexical, Iconic, and Symbolic Gesture Comprehension</i>	
Noriki Nishida, Koji Inoue, Hideki Nakayama, Mayumi Bono and Katsuya Takanashi	514
<i>Fast or Slow? Integrating Fast Intuition and Deliberate Thinking for Enhancing Visual Question Answering</i>	
Songtao Jiang, Chenyi Zhou, Yan Zhang, Yeying Jin and Zuozhu Liu	525
<i>Can Community Notes Replace Professional Fact-Checkers?</i>	
Nadav Borenstein, Greta Warren, Desmond Elliott and Isabelle Augenstein	535
<i>Multilingual Gloss-free Sign Language Translation: Towards Building a Sign Language Foundation Model</i>	
Sihan Tan, Taro Miyazaki and Kazuhiro Nakadai	553
<i>Advancing Sequential Numerical Prediction in Autoregressive Models</i>	
Xiang Fei, Jinghui Lu, Qi Sun, Hao Feng, Yanjie Wang, Wei Shi, An-Lan Wang, Jingqun Tang and Can Huang	562
<i>FEAT: A Preference Feedback Dataset through a Cost-Effective Auto-Generation and Labeling Framework for English AI Tutoring</i>	
Hyein Seo, Taewook Hwang, Yohan Lee and Sangkeun Jung	575
<i>ChronoSense: Exploring Temporal Understanding in Large Language Models with Time Intervals of Events</i>	
Duygu Sezen Islakoglu and Jan-Christoph Kalo	590
<i>Human Alignment: How Much Do We Adapt to LLMs?</i>	
Cazalets Tanguy, Ruben Janssens, Tony Belpaeme and Joni Dambre	603
<i>Dynamic Order Template Prediction for Generative Aspect-Based Sentiment Analysis</i>	
Yonghyun Jun and Hwanhee Lee	614

<i>That doesn't sound right: Evaluating speech transcription quality in field linguistics corpora</i>	
Eric Le Ferrand, Bo Jiang, Joshua Hartshorne and Emily Prud'hommeaux	627
<i>Is That Your Final Answer? Test-Time Scaling Improves Selective Question Answering</i>	
William Jurayj, Jeffrey Cheng and Benjamin Van Durme	636
<i>Acoustic Individual Identification of White-Faced Capuchin Monkeys Using Joint Multi-Species Embeddings</i>	
Álvaro Vega-Hidalgo, Artem Abzaliev, Thore Bergman and Rada Mihalcea	645
<i>SELF-PERCEPT: Introspection Improves Large Language Models' Detection of Multi-Person Mental Manipulation in Conversations</i>	
Danush Khanna, Pratinav Seth, Sidhaarth Sredharan Murali, Aditya Kumar Guru, Siddharth Shukla, Tanuj Tyagi, Sandeep Chaurasia and Kripabandhu Ghosh	660
<i>A Variational Approach for Mitigating Entity Bias in Relation Extraction</i>	
Samuel Mensah, Elena Kochkina, Jabez Magomere, Joy Prakash Sain, Simerjot Kaur and Charese Smiley	676
<i>GenKnowSub: Improving Modularity and Reusability of LLMs through General Knowledge Subtraction</i>	
Mohammadtaha Bagherifard, Sahar Rajabi, Ali Edalat and Yadollah Yaghoobzadeh	685
<i>The Role of Abstract Representations and Observed Preferences in the Ordering of Binomials in Large Language Models</i>	
Zachary Nicholas Houghton, Kenji Sagae and Emily Morgan	695
<i>Can LLMs Understand Unvoiced Speech? Exploring EMG-to-Text Conversion with LLMs</i>	
Payal Mohapatra, Akash Pandey, Xiaoyuan Zhang and Qi Zhu	703
<i>Decoder-Only LLMs can be Masked Auto-Encoders</i>	
Dan Qiao, Yuan Gao, Zheming Yang, Di Yang, Ziheng Wu, Pengcheng Lu, Minghui Qiu, Juntao Li and Min Zhang	713
<i>Mitigating Posterior Saliency Attenuation in Long-Context LLMs with Positional Contrastive Decoding</i>	
Zikai Xiao, Ziyang Wang, Wen MA, Yan Zhang, Wei Shen, WangYan WangYan, Luqi Gong and Zuozhu Liu	724
<i>Sparse-to-Dense: A Free Lunch for Lossless Acceleration of Video Understanding in LLMs</i>	
Xuan Zhang, Cunxiao Du, Sicheng Yu, Jiawei Wu, Fengzhuo Zhang, Wei Gao and Qian Liu	734
<i>Revisiting Uncertainty Quantification Evaluation in Language Models: Spurious Interactions with Response Length Bias Results</i>	
Andrea Santilli, Adam Golinski, Michael Kirchhof, Federico Danieli, Arno Blaas, Miao Xiong, Luca Zappella and Sinead Williamson	743
<i>Memorization Inheritance in Sequence-Level Knowledge Distillation for Neural Machine Translation</i>	
Verna Dankers and Vikas Raunak	760
<i>CoRet: Improved Retriever for Code Editing</i>	
Fabio James Fehr, Prabhu Teja S, Luca Franceschi and Giovanni Zappella	775
<i>Has Machine Translation Evaluation Achieved Human Parity? The Human Reference and the Limits of Progress</i>	
Lorenzo Proietti, Stefano Perrella and Roberto Navigli	790
<i>Diffusion Directed Acyclic Transformer for Non-Autoregressive Machine Translation</i>	
Quan Nguyen-Tri, Cong Dao Tran and Hoang Thanh-Tung	814

<i>Efficient Knowledge Editing via Minimal Precomputation</i>	
Akshat Gupta, Maochuan Lu, Thomas Hartvigsen and Gopala Anumanchipalli	829
<i>Meaning Variation and Data Quality in the Corpus of Founding Era American English</i>	
Dallas Card	841
<i>MindRef: Mimicking Human Memory for Hierarchical Reference Retrieval with Fine-Grained Location Awareness</i>	
Ye Wang, Xinrun Xu and Zhiming Ding	857
<i>LLMs syntactically adapt their language use to their conversational partner</i>	
Florian Kandra, Vera Demberg and Alexander Koller	873
<i>TigerLLM - A Family of Bangla Large Language Models</i>	
Nishat Raihan and Marcos Zampieri	887
<i>From Citations to Criticality: Predicting Legal Decision Influence in the Multilingual Swiss Jurisprudence</i>	
Ronja Stern, Ken Kawamura, Matthias Stürmer, Ilias Chalkidis and Joel Niklaus	897
<i>Revisiting LLMs as Zero-Shot Time Series Forecasters: Small Noise Can Break Large Models</i>	
Junwoo Park, Hyuck Lee, Dohyun Lee, Daehoon Gwak and Jaegul Choo	906
<i>Transferring Textual Preferences to Vision-Language Understanding through Model Merging</i>	
Chen-An Li, Tzu-Han Lin, Yun-Nung Chen and Hung-yi Lee	923
<i>ProgCo: Program Helps Self-Correction of Large Language Models</i>	
Xiaoshuai Song, Yanan Wu, Weixun Wang, Jiaheng Liu, Wenbo Su and Bo Zheng	944
<i>Leveraging Self-Attention for Input-Dependent Soft Prompting in LLMs</i>	
Ananth Muppidi, Abhilash Nandy and Sambaran Bandyopadhyay	960
<i>Inconsistent Tokenizations Cause Language Models to be Perplexed by Japanese Grammar</i>	
Andrew Gambardella, Takeshi Kojima, Yusuke Iwasawa and Yutaka Matsuo	970
<i>Unique Hard Attention: A Tale of Two Sides</i>	
Selim Jerad, Anej Svete, Jiaoda Li and Ryan Cotterell	977
<i>Enhancing Input-Label Mapping in In-Context Learning with Contrastive Decoding</i>	
Keqin Peng, Liang Ding, Yuanxin Ouyang, Meng Fang, Yancheng Yuan and Dacheng Tao . .	997
<i>Different Speech Translation Models Encode and Translate Speaker Gender Differently</i>	
Dennis Fucci, Marco Gaido, Matteo Negri, Luisa Bentivogli, Andre Martins and Giuseppe Attanasio	1005
<i>Rethinking Semantic Parsing for Large Language Models: Enhancing LLM Performance with Semantic Hints</i>	
Kaikai An, Shuzheng Si, Helan Hu, Haozhe Zhao, Yuchi Wang, Qingyan Guo and Baobao Chang	1020
<i>Quantifying Misattribution Unfairness in Authorship Attribution</i>	
Pegah Alipoormolabashi, Ajay Patel and Niranjan Balasubramanian	1030
<i>Zero-Shot Text-to-Speech for Vietnamese</i>	
Thi Vu, Linh The Nguyen and Dat Quoc Nguyen	1042

<i>Can LLMs Generate High-Quality Test Cases for Algorithm Problems? TestCase-Eval: A Systematic Evaluation of Fault Coverage and Exposure</i>	
Zheyuan Yang, Zexi Kuang, Xue Xia and Yilun Zhao	1050
<i>Are Optimal Algorithms Still Optimal? Rethinking Sorting in LLM-Based Pairwise Ranking with Batching and Caching</i>	
Juan Wisznia, Cecilia Bolaños, Juan Tollo, Giovanni Franco Gabriel Marraffini, Agustín Andrés Gianolini, Noe Fabian Hsueh and Luciano Del Corro	1064
<i>TreeCut: A Synthetic Unanswerable Math Word Problem Dataset for LLM Hallucination Evaluation</i>	
Jialin Ouyang	1073
<i>WinSpot: GUI Grounding Benchmark with Multimodal Large Language Models</i>	
Zheng Hui, Yinheng Li, Dan Zhao, Colby Banbury, Tianyi Chen and Kazuhito Koishida	1086
<i>Spurious Correlations and Beyond: Understanding and Mitigating Shortcut Learning in SDOH Extraction with Large Language Models</i>	
Fardin Ahsan Sakib, Ziwei Zhu, Karen Trister Grace, Meliha Yetisgen and Ozlem Uzuner	1097
<i>Enhancing NER by Harnessing Multiple Datasets with Conditional Variational Autoencoders</i>	
Taku Oi and Makoto Miwa	1107
<i>CHEER-Ekman: Fine-grained Embodied Emotion Classification</i>	
Phan Anh Duong, Cat Luong, Divyesh Bommana and Tianyu Jiang	1118
<i>ScanEZ: Integrating Cognitive Models with Self-Supervised Learning for Spatiotemporal Scanpath Prediction</i>	
Ekta Sood, Prajit Dhar, Enrica Troiano, Rosy Southwell and Sidney K. DMello	1132
<i>Improving Fairness of Large Language Models in Multi-document Summarization</i>	
Haoyuan Li, Rui Zhang and Snigdha Chaturvedi	1143
<i>Should I Believe in What Medical AI Says? A Chinese Benchmark for Medication Based on Knowledge and Reasoning</i>	
Yue Wu, Yangmin Huang, Qianyun Du, Lixian Lai, Zhiyang He, Jiaxue Hu and Xiaodong Tao	1155
<i>Rethinking Evaluation Metrics for Grammatical Error Correction: Why Use a Different Evaluation Process than Human?</i>	
Takumi Goto, Yusuke Sakai and Taro Watanabe	1165
<i>Learning Auxiliary Tasks Improves Reference-Free Hallucination Detection in Open-Domain Long-Form Generation</i>	
Chengwei Qin, Wenxuan Zhou, Karthik Abinav Sankararaman, Nanshu Wang, Tengyu Xu, Alexander Radovic, Eryk Helenowski, Arya Talebzadeh, Aditya Tayade, Sinong Wang, Shafiq Joty, Han Fang and Hao Ma	1173
<i>WiCkeD: A Simple Method to Make Multiple Choice Benchmarks More Challenging</i>	
Ahmed Elhady, Eneko Agirre and Mikel Artetxe	1183
<i>Cross-Lingual Representation Alignment Through Contrastive Image-Caption Tuning</i>	
Nathaniel Krasner, Nicholas Lanuzo and Antonios Anastasopoulos	1193
<i>LAMB: A Training-Free Method to Enhance the Long-Context Understanding of SSMs via Attention-Guided Token Filtering</i>	
Zhifan Ye, Zheng Wang, Kejing Xia, Jihoon Hong, Leshu Li, Lexington Whalen, Cheng Wan, Yonggan Fu, Yingyan Celine Lin and Souvik Kundu	1200

<i>Counterfactual-Consistency Prompting for Relative Temporal Understanding in Large Language Models</i>	
Jongho Kim and Seung-won Hwang	1210