

# **Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)**

System Demonstrations

Suzhou, China  
4-9 November 2025

Volume 1 of 2

ISBN: 979-8-3313-2865-8

**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](mailto: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](mailto:curran@proceedings.com)  
Web: [www.proceedings.com](http://www.proceedings.com)

## Table of Contents

<i>Synthetic Data for Evaluation: Supporting LLM-as-a-Judge Workflows with EvalAssist</i> Martín Santillán Cooper, Zahra Ashktorab, Hyo Jin Do, Erik Miehl, Werner Geyer, Jasmina Gajcin, Elizabeth M. Daly, Qian Pan and Michael Desmond . . . . .	1
<i>ROBOTO2: An Interactive System and Dataset for LLM-assisted Clinical Trial Risk of Bias Assessment</i> Anthony Hevia, Sanjana Chintalapati, Veronica Ka Wai Lai, Nguyen Thanh Tam, Wai-Tat Wong, Terry P Klassen and Lucy Lu Wang . . . . .	12
<i>SpiritRAG: A Q&amp;A System for Religion and Spirituality in the United Nations Archive</i> Yingqiang Gao, Fabian Winiger, Patrick Montjourides, Anastassia Shaitarova, Nianlong Gu, Simon Peng-Keller and Gerold Schneider . . . . .	26
<i>LingConv: An Interactive Toolkit for Controlled Paraphrase Generation with Linguistic Attribute Control</i> Mohamed Elgaar and Hadi Amiri . . . . .	42
<i>AgentMaster: A Multi-Agent Conversational Framework Using A2A and MCP Protocols for Multimodal Information Retrieval and Analysis</i> Callie C. Liao, Duoduo Liao and Sai Surya Gadiraju . . . . .	52
<i>The iRead4Skills Intelligent Complexity Analyzer</i> Wafa Aissa, Raquel Amaro, David Antunes, Thibault Bañeras-Roux, Jorge Baptista, Alejandro Catala, Luís Correia, Thomas François, Marcos Garcia, Mario Izquierdo-Álvarez, Nuno Mamede, Vasco Martins, Miguel Neves, Eugénio Ribeiro, Sandra Rodriguez Rey and Elodie Vanzeveren . . . . .	73
<i>AIPOM: Agent-aware Interactive Planning for Multi-Agent Systems</i> Hannah Kim, Kushan Mitra, Chen Shen, Dan Zhang and Estevam Hruschka . . . . .	85
<i>LAD: LoRA-Adapted Diffusion</i> Ruurd Jan Anthonius Kuiper, Lars de Groot, Bram van Es, Maarten van Smeden and Ayoub Bagheri . . . . .	97
<i>Automated Evidence Extraction and Scoring for Corporate Climate Policy Engagement: A Multilingual RAG Approach</i> Imene Kolli, Saeid Vaghefi, Chiara Colesanti Senni, Shantam Raj and Markus Leippold . . . . .	111
<i>GLiNER2: Schema-Driven Multi-Task Learning for Structured Information Extraction</i> Urchade Zaratiana, Gil Pasternak, Oliver Boyd, George Hurn-Maloney and Ash Lewis . . . . .	130
<i>SciClaims: An End-to-End Generative System for Biomedical Claim Analysis</i> Raúl Ortega and Jose Manuel Gomez-Perez . . . . .	141
<i>AgentCPM-GUI: Building Mobile-Use Agents with Reinforcement Fine-Tuning</i> Zhong Zhang, Yaxi Lu, Yikun Fu, Yupeng Huo, Shenzhi Yang, Yesai Wu, Han Si, Xin Cong, Haotian Chen, Yankai Lin, Xie Xie, Wei Zhou, Wang Xu, Zhou Su, Zhongwu Zhai, Xiaoming Liu, Meiyudong, Jianming Xu, Hongyan Tian, Chongyi Wang, Chi Chen, Yuan Yao, Zhiyuan Liu and Maosong Sun . . . . .	155
<i>Marcel: A Lightweight and Open-Source Conversational Agent for University Student Support</i> Jan Trienes, Anastasiia Derzhanskaia, Roland Schwarzkopf, Markus Mühl, Jörg Schlötterer and Christin Seifert . . . . .	181

<i>Alpha-GPT: Human-AI Interactive Alpha Mining for Quantitative Investment</i>	
Saizhuo Wang, Hang Yuan, Leon Zhou, Lionel Ni, Heung-Yeung Shum and Jian Guo . . . . .	196
<i>AgentDiagnose: An Open Toolkit for Diagnosing LLM Agent Trajectories</i>	
Tianyue Ou, Wanyao Guo, Apurva Gandhi, Graham Neubig and Xiang Yue . . . . .	207
<i>Tau-Eval: A Unified Evaluation Framework for Useful and Private Text Anonymization</i>	
Gabriel Loiseau, Damien Sileo, Damien Riquet, Maxime Meyer and Marc Tommasi . . . . .	216
<i>ViDove: A Translation Agent System with Multimodal Context and Memory-Augmented Reasoning</i>	
Yichen Lu, Wei Dai, Jiaen Liu, Ching Wing Kwok, Zongheng Wu, Xudong Xiao, Ao Sun, Sheng fu, Jianyuan Zhan, Yian Wang, Takatomo Saito and Sicheng Lai . . . . .	228
<i>Sanskrit Voyager: Unified Web Platform for Interactive Reading and Linguistic Analysis of Sanskrit Texts</i>	
Giacomo De Luca, Danilo Croce and Roberto Basili . . . . .	244
<i>PromptSuite: A Task-Agnostic Framework for Multi-Prompt Generation</i>	
Eliya Habba, Noam Dahan, Gili Lior and Gabriel Stanovsky . . . . .	254
<i>LionGuard 2: Building Lightweight, Data-Efficient &amp; Localised Multilingual Content Moderators</i>	
Leanne Tan, Gabriel Chua, Ziyu Ge and Roy Ka-Wei Lee . . . . .	264
<i>GraphMind: Interactive Novelty Assessment System for Accelerating Scientific Discovery</i>	
Italo Luis da Silva, Hanqi Yan, Lin Gui and Yulan He . . . . .	286
<i>Pico: A Modular Framework for Hypothesis-Driven Small Language Model Research</i>	
Richard Diehl Martinez, David Demitri Africa, Yuval Weiss, Suchir Salhan, Ryan Daniels and Paula Buttery . . . . .	295
<i>DistaLs: a Comprehensive Collection of Language Distance Measures</i>	
Rob Van Der Goot, Esther Ploeger, Verena Blaschke and Tanja Samardzic . . . . .	307
<i>MedTutor: A Retrieval-Augmented LLM System for Case-Based Medical Education</i>	
Dongsuk Jang, Ziyao Shangguan, Kyle Tegtmeier, Anurag Gupta, Jan T Czerminski, Sophie Chheang and Arman Cohan . . . . .	319
<i>Co-DETECT: Collaborative Discovery of Edge Cases in Text Classification</i>	
Chenfei Xiong, Jingwei Ni, Yu Fan, Vilém Zouhar, Donya Rooein, Lorena Calvo-Bartolomé, Alexander Miserlis Hoyle, Zhijing Jin, Mrinmaya Sachan, Markus Leippold, Dirk Hovy, Mennatallah El-Assady and Elliott Ash . . . . .	354
<i>DVAGen: Dynamic Vocabulary Augmented Generation</i>	
Wei Du, Nuowei Liu, Jie Wang, Jiahao Kuang, Tao Ji, Xiaoling Wang and Yuanbin Wu . . . . .	365
<i>MCPEval: Automatic MCP-based Deep Evaluation for AI Agent Models</i>	
Zhiwei Liu, Jieli Qiu, Shiyu Wang, Jianguo Zhang, Zuxin Liu, Roshan Ram, Haolin Chen, Weiran Yao, Shelby Heinecke, Silvio Savarese, Huan Wang and Caiming Xiong . . . . .	373
<i>SciSketch: An Open-source Framework for Automated Schematic Diagram Generation in Scientific Papers</i>	
Zihang Wang, Yilun Zhao, Kaiyan Zhang, Chen Zhao, Manasi Patwardhan and Arman Cohan	403
<i>MALLM: Multi-Agent Large Language Models Framework</i>	
Jonas Becker, Lars Benedikt Kaesberg, Niklas Bauer, Jan Philip Wahle, Terry Ruas and Bela Gipp	418

<i>SWE-MERA: A Dynamic Benchmark for Agentically Evaluating Large Language Models on Software Engineering Tasks</i>	
Adamenko Pavel, Ivanov Mikhail, Aidar Valeev, Rodion Levichev, Pavel Zadorozhny, Ivan Lopatin, Dmitrii Babaev, Alena Fenogenova and Valentin Malykh . . . . .	440
<i>Open-Theatre: An Open-Source Toolkit for LLM-based Interactive Drama</i>	
Tianyang Xu, Hongqiu Wu, Weiqi Wu and Hai Zhao . . . . .	453
<i>CafGa: Customizing Feature Attributions to Explain Language Models</i>	
Alan David Boyle, Furui Cheng, Vilém Zouhar and Mennatallah El-Assady . . . . .	461
<i>UnityAI Guard: Pioneering Toxicity Detection Across Low-Resource Indian Languages</i>	
Himanshu Beniwal, Reddybathuni Venkat, Rohit Kumar, Birudugadda Srivibhav, Daksh Jain, Pavan Deekshith Doddi, Eshwar Dhande, Adithya Ananth, Kuldeep and Mayank Singh . . . . .	471
<i>BioGraphia: A LLM-Assisted Biological Pathway Graph Annotation Platform</i>	
Xi Xu, Sumin Jo, Adam Officer, Angela Chen, Yufei Huang and Lei Li . . . . .	480
<i>SynthTextEval: Synthetic Text Data Generation and Evaluation for High-Stakes Domains</i>	
Krithika Ramesh, Daniel Smolyak, Zihao Zhao, Nupoor Gandhi, Ritu Agarwal, Margrét V. Bjarnadóttir and Anjalie Field . . . . .	487
<i>Quest2DataAgent: Automating End-to-End Scientific Data Collection</i>	
Tianyu Yang, Yuhan Liu, Sobin Alosious, Ethan A. Brown, Jason R. Rohr, Tengfei Luo and Xiangliang Zhang . . . . .	500
<i>End-to-End Multilingual Automatic Dubbing via Duration-based Translation with Large Language Models</i>	
Hyun-Sik Won, DongJin Jeong, Hyunkyu Choi and Jinwon Kim . . . . .	515
<i>EasyEdit2: An Easy-to-use Steering Framework for Editing Large Language Models</i>	
Ziwen Xu, Shuxun Wang, Kewei Xu, Haoming Xu, Mengru Wang, Xinle Deng, Yunzhi Yao, Guozhou Zheng, Huajun Chen and Ningyu Zhang . . . . .	522
<i>AERA Chat: An Interactive Platform for Automated Explainable Student Answer Assessment</i>	
Jiazheng Li, Artem Bobrov, Runcong Zhao, Cesare Aloisi and Yulan He . . . . .	536
<i>RadEval: A framework for radiology text evaluation</i>	
Justin Xu, Xi Zhang, Javid Abderezaei, Julie Bauml, Roger Boodoo, Fatemeh Haghighi, Ali Ganjizadeh, Eric Brattain, Dave Van Veen, Zaiqiao Meng, David W Eyre and Jean-Benoit Delbrouck . . . . .	546
<i>TinyScientist: An Interactive, Extensible, and Controllable Framework for Building Research Agents</i>	
Haofei Yu, Keyang Xuan, Fenghai Li, Kunlun Zhu, Zijie Lei, Jiaxun Zhang, Ziheng Qi, Kyle Richardson and Jiaxuan You . . . . .	558
<i>KMatrix-2: A Comprehensive Heterogeneous Knowledge Collaborative Enhancement Toolkit for Large Language Model</i>	
Shun Wu, Di Wu, Wangtao Sun, Ziyang Huang, Xiaowei Yuan, Kun Luo, XueYou Zhang, Shizhu He, Jun Zhao and Kang Liu . . . . .	591
<i>GlotEval: A Test Suite for Massively Multilingual Evaluation of Large Language Models</i>	
Hengyu Luo, Zihao Li, Joseph Attieh, Sawal Devkota, Ona de Gibert, Xu Huang, Shaoxiong Ji, Peiqin Lin, Bhavani Sai Praneeth Varma Mantina, Ananda Sreenidhi, Raúl Vázquez, Mengjie Wang, Samea Yusofi, Fei Yuan and Jörg Tiedemann . . . . .	602

<i>MASA: LLM-Driven Multi-Agent Systems for Autoformalization</i>	
Lan Zhang, Marco Valentino and Andre Freitas .....	615
<i>LearnLens: LLM-Enabled Personalised, Curriculum-Grounded Feedback with Educators in the Loop</i>	
Runcong Zhao, Artem Bobrov, Jiazheng Li, Cesare Aloisi and Yulan He .....	625
<i>o-MEGA: Optimized Methods for Explanation Generation and Analysis</i>	
Ľuboš Kriš, Jaroslav Kopčan, Qiwei Peng, Andrej Ridzik, Marcel Veselý and Martin Tamajka	634
<i>EvoAgentX: An Automated Framework for Evolving Agentic Workflows</i>	
Yingxu Wang, Siwei Liu, Jinyuan Fang and Zaiqiao Meng .....	643
<i>OpenRLHF: A Ray-based Easy-to-use, Scalable and High-performance RLHF Framework</i>	
Jian Hu, Xibin Wu, Wei Shen, Jason Klein Liu, Weixun Wang, Songlin Jiang, Haoran Wang, Hao Chen, Bin Chen, Wenkai Fang, Xianyu, Yu Cao, Haotian Xu and Yiming Liu .....	656
<i>ConfReady: A RAG based Assistant and Dataset for Conference Checklist Responses</i>	
Michael Galarnyk, Rutwik Routu, Vidhyakshaya Kannan, Kosha Bheda, Prasun Banerjee, Agam Shah and Sudheer Chava .....	667
<i>TokenSmith: Streamlining Data Editing, Search, and Inspection for Large-Scale Language Model Training and Interpretability</i>	
Mohammad Aflah Khan, Ameya Godbole, Johnny Wei, Ryan Yixiang Wang, James Flemings, Krishna P. Gummadi, Willie Neiswanger and Robin Jia .....	678
<i>LLM×MapReduce-V3: Enabling Interactive In-Depth Survey Generation through a MCP-Driven Hierarchically Modular Agent System</i>	
Yu Chao, Siyu Lin, Xiaorong Wang, Zhu Zhang, Zihan Zhou, Haoyu Wang, Shuo Wang, Jie Zhou, Zhiyuan Liu and Maosong Sun .....	688
<i>GraDeT-HTR: A Resource-Efficient Bengali Handwritten Text Recognition System utilizing Grapheme-based Tokenizer and Decoder-only Transformer</i>	
Md. Mahmudul Hasan, Ahmed Nesar Tahsin Choudhury, Mahmudul Hasan and Md Mosaddek Khan .....	696
<i>AutoIntent: AutoML for Text Classification</i>	
Ilya Alekseev, Roman Solomatin, Darina Rustamova and Denis Kuznetsov .....	707
<i>TruthTorchLM: A Comprehensive Library for Predicting Truthfulness in LLM Outputs</i>	
Duygu Nur Yaldiz, Yavuz Faruk Bakman, Sungmin Kang, Alperen Öziş, Hayrettin Eren Yildiz, Mitash Ashish Shah, Zhiqi Huang, Anoop Kumar, Alf Samuel, Daben Liu, Sai Praneeth Karimireddy and Salman Avestimehr .....	717
<i>The Dangers of Indirect Prompt Injection Attacks on LLM-based Autonomous Web Navigation Agents: A Demonstration</i>	
Sam Johnson, Viet Pham and Thai Le .....	729
<i>LaTeXMT: Machine Translation for LaTeX Documents</i>	
Calvin Hoy, Samuel Frontull and Georg Moser .....	739
<i>LangVAE and LangSpace: Building and Probing for Language Model VAEs</i>	
Danilo Carvalho, Yingji Zhang, Harriet Unsworth and Andre Freitas .....	749
<i>PresentAgent: Multimodal Agent for Presentation Video Generation</i>	
Jingwei Shi, Zeyu Zhang, Biao Wu, Yanjie Liang, Meng Fang, Ling Chen and Yang Zhao ...	760

<i>PromptSculptor: Multi-Agent Based Text-to-Image Prompt Optimization</i> Dawei Xiang, Wenyan Xu, Kexin Chu, Tianqi Ding, Zixu Shen, Yiming Zeng, Jianchang Su and Wei Zhang .....	774
<i>EasyDistill: A Comprehensive Toolkit for Effective Knowledge Distillation of Large Language Models</i> Chengyu Wang, Junbing Yan, Wenrui Cai, Yuanhao Yue and Jun Huang .....	787
<i>AM4DSP: Argumentation Mining in Structured Decentralized Discussion Platforms for Deliberative Democracy</i> Sofiane Elguendouze, Lucas Anastasiou, Erwan Hain, Elena Cabrio, Anna De Liddo and Serena Villata .....	796
<i>TRACE: Training and Inference-Time Interpretability Analysis for Language Models</i> Nura Aljaafari, Danilo Carvalho and Andre Freitas .....	806
<i>MathBuddy: A Multimodal System for Affective Math Tutoring</i> Debanjana Kar, Leopold Böss, Dacia Braca, Sebastian Maximilian Dennerlein, Nina Christine Hubig, Philipp Wintersberger and Yufang Hou .....	821
<i>PledgeTracker: A System for Monitoring the Fulfilment of Pledges</i> Yulong Chen, Michael Sejr Schlichtkrull, Zhenyun Deng, David Corney, Nasim Asl, Joshua Salisbury, Andrew Dudfield and Andreas Vlachos .....	839
<i>Interactive Training: Feedback-Driven Neural Network Optimization</i> Wentao Zhang, Yang Young Lu and Yuntian Deng .....	851
<i>Metamo: Empowering Large Language Models with Psychological Distortion Detection for Cognition-aware Coaching</i> Hajime Hotta, Huu-Loi Le, Manh-Cuong Phan and Minh-Tien Nguyen .....	862
<i>InTriage: Intelligent Telephone Triage in Pre-Hospital Emergency Care</i> Kai He, Qika Lin, Hao Fei, Eng Siong Chng, Dehan Hong, Marcus Eng Hock Ong and Mengling Feng .....	873
<i>From Behavioral Performance to Internal Competence: Interpreting Vision-Language Models with VLM-Lens</i> Hala Sheta, Eric Haoran Huang, Shuyu Wu, Ilija Alenabi, Jiajun Hong, Ryker Lin, Ruoxi Ning, Daniel Wei, Jialin Yang, Jiawei Zhou, Ziqiao Ma and Freda Shi .....	886
<i>ResearStudio: A Human-intervenable Framework for Building Controllable Deep Research Agents</i> Linyi Yang and Yixuan Weng .....	896
<i>OpenS2S: Advancing Fully Open-Source End-to-End Empathetic Large Speech Language Model</i> Chen Wang, Tianyu Peng, Wen Yang, YiNan Bai, Guangfu Wang, Jun Lin, Lanpeng Jia, Lingxiang Wu, Jinqiao Wang, Chengqing Zong and Jiajun Zhang .....	906
<i>PDFMathTranslate: Scientific Document Translation Preserving Layouts</i> Rongxin Ouyang, Chang Chu, Zhikuan Xin and Xiangyao Ma .....	918
<i>CrowdAgent: Multi-Agent Managed Multi-Source Annotation System</i> Maosheng Qin, Renyu Zhu, Mingxuan Xia, Chenchenkai, Zhen Zhu, Minmin Lin, Junbo Zhao, Lu Xu, Changjie Fan, Runze Wu and Haobo Wang .....	925
<i>Bratly: A Python Extension for BRAT Functionalities</i> Jamil Zahir, Jean-Philippe Goldman, Nikola Bjelogrić, Mina Bjelogrić and Christian Lovis	943

<i>BAREC Demo: Resources and Tools for Sentence-level Arabic Readability Assessment</i>	
Kinda Altarbouch, Khalid N. Elmadani, Ossama Obeid, Hanada Taha and Nizar Habash . . . .	950
<i>Easy Dataset: A Unified and Extensible Framework for Synthesizing LLM Fine-Tuning Data from Unstructured Documents</i>	
Ziyang Miao, Qiyu Sun, Jingyuan Wang, Yuchen Gong, Yaowei Zheng, Shiqi Li and Richong Zhang . . . . .	960
<i>SlackAgents: Scalable Collaboration of AI Agents in Workspaces</i>	
Zhiwei Liu, Weiran Yao, Zuxin Liu, Juntao Tan, Jianguo Zhang, Frank Wang, Sukhandeep Nahal, Huan Wang, Shelby Heinecke, Silvio Savarese and Caiming Xiong . . . . .	969
<i>Open Political Corpora: Structuring, Searching, and Analyzing Political Text Collections with PoliCorp</i>	
Nina Smirnova, Muhammad Ahsan Shahid and Philipp Mayr . . . . .	983