

EMNLP 2023

**The 2023 Conference on Empirical Methods in Natural
Language Processing**

Proceedings of the System Demonstrations

December 6-10, 2023

©2023 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-8-89176-067-7

Introduction

Welcome to the proceedings of the system demonstration track of the 2023 Conference on Empirical Methods in Natural Language Processing on December 6th – December 10th, 2023. For the EMNLP 2023 system demonstration track, we received a record number of 203 submissions, of which 52 were selected for inclusion in the program (acceptance rate of 25%) after being reviewed by at least three members of the program committee, while a small number of papers received only two reviews. We would like to thank the members of the program committee for their timely help in reviewing the submissions. Lastly, we thank the many authors that submitted their work to the demonstrations track. This year, the EMNLP conference is a hybrid event. The demonstration papers will be presented through pre-recorded talks and in presence during the poster sessions.

Yansong Feng and Els Lefever
EMNLP 2023 System Demonstration Chairs

Organizing Committee

Demonstration Chairs

Yansong Feng, Peking University, China

Els Lefever, LT3, Ghent University, Belgium

Program Committee

Best Demo Paper Award Committee

Walter Daelemans, University of Antwerp
Yusuke Miyao, the University of Tokyo
Djamé Seddah, Inria Paris

Program Committee

Alan Akbik, Humboldt-Universität zu Berlin
Ekaterina Artemova, Toloka.AI
Giosuè Baggio, Norwegian University of Science and Technology
Ayoub Bagheri, Department of Methodology and Statistics, Utrecht University
Timothy Baldwin, MBZUAI
Denilson Barbosa, University of Alberta
Farah Benamara, University of toulouse
Gábor Berend, University Of Szeged
Marianna Bolognesi, University of Bologna
Francis Bond, Palacký University
Georgeta Bordea, Université de Bordeaux
Johan Bos, University of Groningen
Elena Cabrio, Université Côte d'Azur, Inria, CNRS, I3S
Yi Cai, South China University of Technology
Deng Cai, The Chinese University of Hong Kong
Jose Camacho-Collados, Cardiff University
Hejing Cao, Peking University
Tommaso Caselli, Rijksuniversiteit Groningen
Yekun Chai, Baidu
Franklin Chang, Kobe City University of Foreign Studies
Wanxiang Che, Harbin Institute of Technology
Chung-Chi Chen, National Institute of Advanced Industrial Science and Technology
Guanyi Chen, Central China Normal University
Kehai Chen, School of Computer Science and Technology, Harbin Institute of Technology
Sanyuan Chen, Harbin Institute of Technology
Xinchi Chen, Amazon AWS
Yubo Chen, Institute of Automation, Chinese Academy of Sciences
Prafulla Kumar Choubey, Salesforce AI Research
Bonaventura Coppola, University of Trento
Heriberto Cuayahuitl, University of Lincoln
Luna De Bruyne, CLiPS, University of Antwerp
Orphee De Clercq, LT3, Ghent University
Xiang Deng, The Ohio State University
Luigi Di Caro, University of Turin
Jakub Dotlacil, Utrecht University
Zhicheng Dou, Renmin University of China
Zi-Yi Dou, UCLA
Aleksandr Drozd, RIKEN Center for Computational Science
Haowei Du, Peking University
Steffen Eger, NLLG Lab, Bielefeld University

Stefano Faralli, University of Rome Sapienza
Jiazhan Feng, Peking University
Siva Reddy Gangireddy, SoapBox Labs
Voula Giouli, Institute for Language & Speech Processing, ATHENA Research & Innovation Centre
Udo Hahn, Friedrich-Schiller-Universitaet Jena
Samar Haider, University of Pennsylvania
Lifeng Han, The University of Manchester
Wenjuan Han, Beijing Jiaotong University
Sanda Harabagiu, University of Texas at Dallas
Iris Hendrickx, Centre for Language Studies, Radboud University Nijmegen
Delia Irazu Hernandez Farias, Instituto Nacional de Astrofisica, Optica y Electronica
Xudong Hong, Saarland University / MPI Informatics
Veronique Hoste, LT3, Ghent University
Baotian Hu, Harbin Institue of Technology, Shenzhen
Yutong Hu, Peking University
Xinyu Hua, Bloomberg
Quzhe Huang, Peking University
Shujian Huang, National Key Laboratory for Novel Software Technology, Nanjing University
Nancy Ide, Vassar College/Brandeis University
Joseph Marvin Imperial, University of Bath
Radu Tudor Ionescu, University of Bucharest
Yichen Jiang, University of North Carolina at Chapel Hill
Elena Karajosova, Freie Universität
Mladen Karan, Queen Mary University
Yoshihide Kato, Nagoya University
Yova Kementchedjhieva, University of Copenhagen
Salam Khalifa, Stony Brook University
Halil Kilicoglu, University of Illinois at Urbana-Champaign
Dong-Jin Kim, Hanyang University
Mamoru Komachi, Hitotsubashi University
Rik Koncel-Kedziorski, Kensho Technologies
Maarit Koponen, University of Eastern Finland
Viet Lai, University of Oregon
Yuxuan Lai, the Open University of China
Kyungho Lee, SOCO AI
Manling Li, Northwestern University
Maoxi Li, School of Computer Information Engineering, Jiangxi Normal University
Xintong Li, Apple
Zuchao Li, Wuhan University
Jiangming Liu, Yunnan University
Tianyu Liu, Peking University
Xiao Liu, Microsoft Research Asia
Yang Janet Liu, Georgetown University
Aaron Maladry, Ghent University
Alda Mari, <http://www.institutnicod.org/>
Ilia Markov, Vrije Universiteit Amsterdam, CLTL
Eugenio Martínez Cámara, University of Jaén
John P. McCrae, Insight Center for Data Analytics, National University of Ireland Galway
Pushkar Mishra, Meta AI
Amita Misra, Amazon

Richard Moot, CNRS
Gaku Morio, Research & Development Group, Hitachi America, Ltd.
Huy Nguyen, Amazon
Yixin Nie, Meta
Robert Östling, Department of Linguistics, Stockholm University
Jessica Ouyang, University of Texas at Dallas
Alexander Panchenko, Skolkovo Institute of Science and Technology
Jiaxin Pei, University of Michigan
Massimo Poesio, Queen Mary University of London
Octavian Popescu, IBM Yorktown Research
Adithya Pratapa, Carnegie Mellon University
Anil Ramakrishna, Amazon
Julia Rayz, Purdue University
Steffen Remus, Hamburg University
Alexander Rush, Cornell University
Magnus Sahlgren, AI Sweden
Shuming Shi, Tencent AI Lab
Lei Shu, Google Research
Jyotika Singh, Placemakr
Pranaydeep Singh, LT3, University of Ghent
Ionut-Teodor Sorodoc, Amazon
Mark Steedman, University of Edinburgh
Carlo Strapparava, FBK-irst
Lixin Su, Institute of Computing Technology, Chinese Academy of Sciences
Chengjie Sun, Harbin Institute of Technology
Kai Sun, Meta
Marek Šuppa, Comenius University in Bratislava
Colin Swaelens, Ghent University
Aarne Talman, University of Helsinki
Chongyang Tao, Microsoft Corporation
Mingxu Tao, Peking University
Irina Temnikova, Big Data for Smart Society Institute (GATE), Bulgaria
Arda Tezcan, LT3, Language and Translation Technology Team, Ghent University
Christoph Tillmann, IBM Research
Marco Turchi, Zoom Video Communications
Rik van Noord, University of Groningen
Bram Vanroy, KU Leuven
Di Wang, ContextLogic Inc
Ke Wang, Huawei Technologies Ltd.
Qingyun Wang, University of Illinois at Urbana-Champaign
Xun Wang, Microsoft
Wei Wang, Apple AI/ML
Martijn Wieling, University of Groningen
Yang Wu, Harbin Institute of Technology
Zirui Wu, Peking University
Jiarong Xu, Fudan University
Jingjing Xu, Shanghai AI Lab
Ruo Chen Xu, Microsoft
Silei Xu, Stanford University
Wei Xu, Georgia Institute of Technology
Yan Xu, Hong Kong University of Science and Technology

Jingfeng Yang, Amazon
Yaqin Yang, Paypal
Shuang (Sophie) Zhai, University of Oklahoma
Biao Zhang, Google Research
Chen Zhang, Peking University
Meishan Zhang, Harbin Institute of Technology (Shenzhen), China
Tianlin Zhang, The University of Manchester
Wen Zhang, Zhejiang University
Jie Zhao, Microsoft
Xiang Zhao, National University of Defense Technology
Ming Zhong, University of Illinois Urbana-Champaign
Tiantian Zhu, Harbin Institute of Technology (Shenzhen)

Table of Contents

<i>Fabricator: An Open Source Toolkit for Generating Labeled Training Data with Teacher LLMs</i> Jonas Golde, Patrick Haller, Felix Hamborg, Julian Risch and Alan Akbik	1
<i>End-to-End Evaluation for Low-Latency Simultaneous Speech Translation</i> Christian Huber, Tu Anh Dinh, Carlos Mullov, Ngoc-Quan Pham, Thai Binh Nguyen, Fabian Retkowsky, Stefan Constantin, Enes Ugan, Danni Liu, Zhaolin Li, Sai Koneru, Jan Niehues and Alexander Waibel	12
<i>CHATREPORT: Democratizing Sustainability Disclosure Analysis through LLM-based Tools</i> Jingwei Ni, Julia Bingler, Chiara Colesanti-Senni, Mathias Kraus, Glen Gostlow, Tobias Schimanski, Dominik Stambach, Saeid Ashraf Vaghefi, Qian Wang, Nicolas Webersinke, Tobias Wekhof, Tingyu Yu and Markus Leippold	21
<i>RaLLe: A Framework for Developing and Evaluating Retrieval-Augmented Large Language Models</i> Yasuto Hoshi, Daisuke Miyashita, Youyang Ng, Kento Tatsuno, Yasuhiro Morioka, Osamu Torii and Jun Deguchi	52
<i>VIST5: An Adaptive, Retrieval-Augmented Language Model for Visualization-oriented Dialog</i> Henrik Voigt, Nuno Carvalhais, Monique Meuschke, Markus Reichstein, Sina Zarrie and Kai Lawonn	70
<i>H2O Open Ecosystem for State-of-the-art Large Language Models</i> Arno Candell, Jon McKinney, Philipp Singer, Pascal Pfeiffer, Maximilian Jeblick, Chun Ming Lee and Marcos Conde	82
<i>Koala: An Index for Quantifying Overlaps with Pre-training Corpora</i> Thuy-Trang Vu, Xuanli He, Gholamreza Haffari and Ehsan Shareghi	90
<i>Sudowoodo: A Chinese Lyric Imitation System with Source Lyrics</i> Yongzhu Chang, Rongsheng Zhang, Lin Jiang, Qihang Chen, Le Zhang and Jiashu Pu	99
<i>ConvLab-3: A Flexible Dialogue System Toolkit Based on a Unified Data Format</i> Qi Zhu, Christian Geishauser, Hsien-chin Lin, Carel van Niekerk, Baolin Peng, Zheng Zhang, Shutong Feng, Michael Heck, Nurul Lubis, Dazhen Wan, Xiaochen Zhu, Jianfeng Gao, Milica Gasic and Minlie Huang	106
<i>FLEEK: Factual Error Detection and Correction with Evidence Retrieved from External Knowledge</i> Farima Fatahi Bayat, Kun Qian, Benjamin Han, Yisi Sang, Anton Belyy, Samira Khorshidi, Fei Wu, Ihab Ilyas and Yunyao Li	124
<i>YATO: Yet Another deep learning based Text analysis Open toolkit</i> Zeqiang Wang, Yile Wang, Jiageng Wu, Zhiyang Teng and Jie Yang	131
<i>Spacerini: Plug-and-play Search Engines with Pyserini and Hugging Face</i> Christopher Akiki, Odunayo Ogundepo, Aleksandra Piktus, Xinyu Zhang, Akintunde Oladipo, Jimmy Lin and Martin Potthast	140
<i>Adapters: A Unified Library for Parameter-Efficient and Modular Transfer Learning</i> Clifton Poth, Hannah Sterz, Indraneil Paul, Sukannya Purkayastha, Leon Engländer, Timo Imhof, Ivan Vuli, Sebastian Ruder, Iryna Gurevych and Jonas Pfeiffer	149
<i>INTELMO: Enhancing Models' Adoption of Interactive Interfaces</i> Chunxu Yang, Chien-Sheng Wu, Lidiya Murakhovska, Philippe Laban and Xiang Anthony Chen	161

<i>Humanoid Agents: Platform for Simulating Human-like Generative Agents</i> Zhilin Wang, Yu Ying Chiu and Yu Cheung Chiu	167
<i>TP-Detector: Detecting Turning Points in the Engineering Process of Large-scale Projects</i> Qi Wu, WenHan Chao, Xian Zhou and Zhunchen Luo	177
<i>CLEVA: Chinese Language Models EVALuation Platform</i> Yanyang Li, Jianqiao Zhao, Duo Zheng, Zi-Yuan Hu, Zhi Chen, Xiaohui Su, Yongfeng Huang, Shijia Huang, Dahua Lin, Michael Lyu and Liwei Wang	186
<i>DOPA METER – A Tool Suite for Metrical Document Profiling and Aggregation</i> Christina Lohr and Udo Hahn	218
<i>Muted: Multilingual Targeted Offensive Speech Identification and Visualization</i> Christoph Tillmann, Aashka Trivedi, Sara Rosenthal, Santosh Borse, Rong Zhang, Avirup Sil and Bishwaranjan Bhattacharjee	229
<i>Gentopia.AI: A Collaborative Platform for Tool-Augmented LLMs</i> Binfeng Xu, Xukun Liu, Hua Shen, Zeyu Han, Yuhan Li, Murong Yue, Zhiyuan Peng, Yuchen Liu, Ziyu Yao and Dongkuan Xu	237
<i>MusicAgent: An AI Agent for Music Understanding and Generation with Large Language Models</i> Dingyao Yu, Kaitao Song, Peiling Lu, Tianyu He, Xu Tan, Wei Ye, Shikun Zhang and Jiang Bian 246	
<i>SentAlign: Accurate and Scalable Sentence Alignment</i> Steinthor Steingrímsson, Hrafn Loftsson and Andy Way	256
<i>QACheck: A Demonstration System for Question-Guided Multi-Hop Fact-Checking</i> Liangming Pan, Xinyuan Lu, Min-Yen Kan and Preslav Nakov	264
<i>RobustQA: A Framework for Adversarial Text Generation Analysis on Question Answering Systems</i> Yasaman Boreshban, Seyed Morteza Mirbostani, Seyedeh Fatemeh Ahmadi, Gita Shojaei, Fate- meh Kamani, Gholamreza Ghassem-Sani and Seyed Abolghasem Mirroshandel	274
<i>Kandinsky: An Improved Text-to-Image Synthesis with Image Prior and Latent Diffusion</i> Anton Razzhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov, Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov and Denis Dimitrov	286
<i>NewsRecLib: A PyTorch-Lightning Library for Neural News Recommendation</i> Andreea Iana, Goran Glavaš and Heiko Paulheim	296
<i>MiniChain: A Small Library for Coding with Large Language Models</i> Alexander Rush	311
<i>Okapi: Instruction-tuned Large Language Models in Multiple Languages with Reinforcement Learning from Human Feedback</i> Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi and Thien Nguyen	318
<i>SAGEViz: Schema GEneration and Visualization</i> Sugam Devare, Mahnaz Koupaee, Gautham Gunapati, Sayontan Ghosh, Sai Vallurupalli, Yash Kumar Lal, Francis Ferraro, Nathanael Chambers, Greg Durrett, Raymond Mooney, Katrin Erk and Niranjan Balasubramanian	328
<i>Thresh: A Unified, Customizable and Deployable Platform for Fine-Grained Text Evaluation</i> David Heineman, Yao Dou and Wei Xu	336

<i>InsightPilot: An LLM-Empowered Automated Data Exploration System</i> Pingchuan Ma, Rui Ding, Shuai Wang, Shi Han and Dongmei Zhang	346
<i>SynJax: Structured Probability Distributions for JAX</i> Miloš Stanojević and Laurent Sartran	353
<i>RESIN-EDITOR: A Schema-guided Hierarchical Event Graph Visualizer and Editor</i> Khanh Duy Nguyen, Zixuan Zhang, Reece Suchocki, Sha Li, Martha Palmer, Susan Windisch Brown, Jiawei Han and Heng Ji	365
<i>DRGCoder: Explainable Clinical Coding for the Early Prediction of Diagnostic-Related Groups</i> Daniel Hajjaligol, Derek Kaknes, Tanner Barbour, Daphne Yao, Chris North, Jimeng Sun, David Liem and Xuan Wang	373
<i>CAMRA: Copilot for AMR Annotation</i> Jon Cai, Shafiuddin Rehan Ahmed, Julia Bonn, Kristin Wright-Bettner, Martha Palmer and James H. Martin	381
<i>Reaction Miner: An Integrated System for Chemical Reaction Extraction from Textual Data</i> Ming Zhong, Siru Ouyang, Yizhu Jiao, Priyanka Kargupta, Leo Luo, Yanzhen Shen, Bobby Zhou, Xianrui Zhong, Xuan Liu, Hongxiang Li, Jinfeng Xiao, Minhao Jiang, Vivian Hu, Xuan Wang, Heng Ji, Martin Burke, Huimin Zhao and Jiawei Han	389
<i>CHAMP: Efficient Annotation and Consolidation of Cluster Hierarchies</i> Arie Cattan, Tom Hope, Doug Downey, Roy Bar-Haim, Lilach Eden, Yoav Kantor and Ido Dagan	403
<i>Prompt2Model: Generating Deployable Models from Natural Language Instructions</i> Vijay Viswanathan, Chenyang Zhao, Amanda Bertsch, Tongshuang Wu and Graham Neubig	413
<i>NewsSense: Reference-free Verification via Cross-document Comparison</i> Jeremiah Milbauer, Ziqi Ding, Zhijin Wu and Tongshuang Wu	422
<i>NeMo Guardrails: A Toolkit for Controllable and Safe LLM Applications with Programmable Rails</i> Traian Rebedea, Razvan Dinu, Makesh Narsimhan Sreedhar, Christopher Parisien and Jonathan Cohen	431
<i>LM-Polygraph: Uncertainty Estimation for Language Models</i> Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin and Artem Shelmanov	446
<i>Descriptive Knowledge Graph in Biomedical Domain</i> Kerui Zhu, Jie Huang and Kevin Chen-Chuan Chang	462
<i>Prompterator: Iterate Efficiently towards More Effective Prompts</i> Samuel Sučík, Daniel Skala, Andrej Švec, Peter Hraška and Marek Šuppa	471
<i>ZhuJiu: A Multi-dimensional, Multi-faceted Chinese Benchmark for Large Language Models</i> Baoli Zhang, Haining Xie, Pengfan Du, Junhao Chen, Pengfei Cao, Yubo Chen, Shengping Liu, Kang Liu and Jun Zhao	479
<i>PaperMage: A Unified Toolkit for Processing, Representing, and Manipulating Visually-Rich Scientific Documents</i> Kyle Lo, Zejiang Shen, Benjamin Newman, Joseph Chee Chang, Russell Authur, Erin Bransom, Stefan Candra, Yoganand Chandrasekhar, Regan Huff, Bailey Kuehl, Amanpreet Singh, Chris Wilhelm, Angele Zamarron, Marti A. Hearst, Daniel Weld, Doug Downey and Luca Soldaini	495

<i>OmniEvent: A Comprehensive, Fair, and Easy-to-Use Toolkit for Event Understanding</i> Hao Peng, Xiaozhi Wang, Feng Yao, Zimu Wang, Chuzhao Zhu, Kaisheng Zeng, Lei Hou and Juanzi Li	508
<i>CocoSciSum: A Scientific Summarization Toolkit with Compositional Controllability</i> Yixi Ding, Yanxia Qin, Qian Liu and Min-Yen Kan	518
<i>CoLLiE: Collaborative Training of Large Language Models in an Efficient Way</i> Kai Lv, Shuo Zhang, Tianle Gu, Shuhao Xing, Jiawei Hong, Keyu Chen, Xiaoran Liu, Yuqing Yang, Honglin Guo, Tengxiao Liu, Yu Sun, Qipeng Guo, Hang Yan and Xipeng Qiu	527
<i>Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding</i> Hang Zhang, Xin Li and Lidong Bing	543
<i>SummHelper: Collaborative Human-Computer Summarization</i> Aviv Slobodkin, Niv Nachum, Shmuel Amar, Ori Shapira and Ido Dagan	554
<i>ModelScope-Agent: Building Your Customizable Agent System with Open-source Large Language Models</i> Chenliang Li, He Chen, Ming Yan, Weizhou Shen, Haiyang Xu, Zhikai Wu, Zhicheng Zhang, Wenmeng Zhou, Yingda Chen, Chen Cheng, Hongzhu Shi, Ji Zhang, Fei Huang and Jingren Zhou	566
<i>EfficientOCR: An Extensible, Open-Source Package for Efficiently Digitizing World Knowledge</i> Tom Bryan, Jacob Carlson, Abhishek Arora and Melissa Dell	579