ACL 2026

BioNLP 2026 and Shared Tasks

Proceedings of the 25th Workshop on Biomedical Language Processing

July 3-4, 2026

©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 979-8-89176-434-7

BioNLP 2026, a year in review: Asking LLMs for the facts, all the facts, and nothing but the facts

Dina Demner-Fushman, Sophia Ananiadou, Kirk Roberts, Junichi Tsujii

BioNLP 2026 marks a quarter-century of the efforts to bring together computational linguists whose work is dedicated to improving health through language technologies, ranging from foundational tasks to clinical applications. Large Language Models (LLMs) continue to be the mainstay of Biomedical Language Processing, while the scope of BioNLP research continues to expand across foundational tasks, applications, languages and modalities. In 2026, we see increasing efforts to integrate textual features with visual and sequencing data; new approaches to named entity recognition and linking; work in several languages other than English; and applications ranging from molecular biology to clinical studies. Complex language technology tasks, such as question answering and summarization, as well as data generation and text mining are also strongly represented. Concerns about potential harms and irresponsible use of AI applications are being addressed through growing research into evaluation, debiasing, and understanding of models' behavior. The submissions to the BioNLP 2026 workshop and the Shared Tasks demonstrated once again that the workshop– sponsored by the ACL Special Interest Group on Biomedical Natural Language Processing (SIGBioMed)– is the preferred venue for the groundbreaking research and applications in Biomedical Language Processing, which encompasses biological, clinical and non-health professional sub-languages, among others. BioNLP remains the flagship venue and the broadest forum in biomedical language processing, accepting high-quality research across all tasks, methods and languages. The quality and diversity of submissions continues to impress the program committee and the organizers.

BioNLP 2026 received 109 submissions spanning foundational research, biomedical language processing, clinical applications, and development of new datasets and benchmarks. Notably, in addition to the shared tasks that focused on factuality and accuracy of the results generated by large language models, many workshop submissions take a closer look at the nature of hallucinations and approaches to ascertain the factuality of the results generated by the models for various tasks. Moreover, there is a renewed interest in improving the quality of the existing benchmarks and designing new evaluation approaches that not only evaluate factuality, but also require grounding and explanations of the process that generated the results.

BioNLP has a long-standing tradition of sponsoring Shared Tasks. In 2026, we invited SIGBioMed members to submit a description of a shared task to be included with the BioNLP proposal. We received four strong detailed descriptions of the tasks, which were reviewed by the workshop organizers. These well-defined and timely tasks collocated with BioNLP 2026 are briefly described below.

MedExACT task involved detection and labeling of medical decisions in ICU discharge summaries, with evaluation metrics emphasizing both accuracy and fairness across demographic and disease subgroups at the span and token levels, as well as through stratified analyses to measure robustness against biases in sex, race, English proficiency, and disease type. Baseline models such as RoBERTa indicated the complexity of the task. The participants were supported with expedited access to MedDec through PhysioNet, a public leaderboard, and a starter kit in Python.

PsyDefDetect: Detecting Psychological Defense Mechanisms in Conversations. This task focused on classifying Seeker utterances in supportive conversations into specific Psychological Defense Levels based on the Defense Mechanism Rating Scales (DMRS) framework. The benchmark addresses the challenge of capturing subtle linguistic cues of deep-seated psychological mechanisms within highly informal and context-dependent emotional dialogues. This initiative supports research at the intersection of clinical psychology and NLP, aiming to operationalize complex psychological constructs for computational

analysis.

MedGenVidQA task focused on grounding answers with reference attribution to mitigate generation of false statements by LLMs when answering biomedical questions. MedGenVidQA 2026 introduces generation of multimodal answers from textual and visual sources with citations, leveraging PubMed and HealthVidQA as multimodal sources. The test set is based on the information requests submitted by self-identified non-clinicians to the MedlinePlus service provided by the National Library of Medicine. The evaluation leverages BioACE, an automated evaluation framework that strongly correlates with human evaluation on the BioGen 2024 textual dataset.

ClinicalSkillQA formulated clinical skill understanding and continuous perception for clinical skill assessment as an ordering task: the MLLM is required to arrange shuffled key frames into a coherent sequence of clinical actions and to provide explanations for the resulting order. The dataset is constructed from video clips of medical student clinical procedures, collected from Zhongnan Hospital of Wuhan University and Cofun.

The overviews of the tasks are included in the workshop program. The participants in all Shared Tasks present their work on July 4th.

The keynote by Annika Marie Schoene, PhD, Assistant Professor, Bouve College of Health Sciences, is titled: AI Safety in Healthcare: Ethical and Technical Considerations.
The rapid integration of artificial intelligence into healthcare, spanning ambient documentation, virtual nursing, and medical imaging, has outpaced the development of robust oversight mechanisms. While the global AI-in-healthcare market is projected to exceed 868 billion by 2030, incidents of algorithmic bias and unequal care delivery have exposed significant gaps between responsible AI principles and clinical implementation. This talk examines AI safety as a foundational pillar of responsible AI in health contexts, arguing that fairness, transparency, accountability, and human autonomy cannot remain aspirational without corresponding technical and governance infrastructure. I survey existing frameworks and identify a critical missing layer of concrete, operationalizable evaluation methods that bridge ethical principles and technical practice. To address this gap, I present work toward an actionable framework for responsible AI integration in clinical settings, grounded in the AI Ethics Box, a structured taxonomy derived from biomedical ethics. The framework maps ethical domains to specific technical tests, supporting both pre- and post-deployment evaluation. I also describe an ongoing co-design process with clinical and community stakeholders to ensure real-world feasibility. I conclude that AI in healthcare demands genuinely interdisciplinary research to build tools that improve health outcomes without compromising patient safety.

Biography: Annika Marie Schoene, PhD is a computer scientist and researcher in AI safety, working on the evaluation, robustness, and security of large-scale AI systems, including large language models. She develops technical methods and evaluation frameworks to identify and mitigate high-risk behaviors such as jailbreaks, harmful outputs, and unsafe system behavior, with the goal of enabling the safe and trustworthy deployment of AI in public health, health systems, and healthcare settings. She is currently an Assistant Professor in the Department of Public Health and Health Sciences and Technical Lead for the Responsible AI Practice at Northeastern University. Beyond her core research, she works across academic, industry, public-sector and not-for-profit settings to generate evidence that supports non-technical stakeholders and policymakers in making informed decisions that reduce algorithmic harm and inform organizational decision-making. She is also a Visiting Scientist at MaineHealth and the University of Southampton (UK), and a Faculty Fellow at the Institute for Social Justice and Healthy Equity, and serves as a Scientific Expert Advisor at Meta on AI safety. She holds a PhD in Computer Science from the University of Hull (UK). During her doctoral training, she conducted research on machine and deep learning methods for analyzing complex real-world text and interned at IBM Research (UK), continuing this collaboration throughout her PhD and postdoctoral work. She completed her postdoctoral training at the University of Manchester's National Centre for Text Mining (NaCTeM), where she worked on natural language processing in health-related contexts. Prior to her current role, she was a Research Scientist

at the Institute for Experiential AI (EAI), working primarily on research involving the development and evaluation of AI methods in applied health contexts.

As always, we are deeply grateful to the authors of the submitted papers and to the reviewers (listed elsewhere in this volume) who produced three thorough and thoughtful reviews for each paper in a very short review period. The quality of submitted work continues to grow, and the organizers are truly grateful to the members of our Program Committee, who helped us to determine which work was ready to be presented, and which would benefit from the additional experiments and analyses suggested by the reviewers. As in years past, we are looking forward to a productive workshop and hoping it will foster new collaborations and research. This will enable our community to continue making valuable contributions to public health and well-being, as well as to basic and clinical research.

Program Committee

Chairs

Sophia Ananiadou, University of Manchester
Dina Demner-Fushman, National Library of Medicine
Kirk Roberts, University of Texas Health Science Center at Houston

Program Committee

Abdulrahman Aal Abdulsalam, Sultan Qaboos University
Rohit Agarwal, UiT The Arcitic University of Norway
Aizierjiang Aiersilan, The George Washington University
Ebrahim Alharbi, The University of Sheffield
Daniel Andrade, Hiroshima University
Eiji Aramaki, NAIST, Japan
Niloofar Arazkhani, University of Pittsburgh
Steven Au, UCSC
Davis Bartels, National Library of Medicine
Hadas Ben Atya, Technion
Krushil Bhojani, Suny Polytechnic Institute
Madeline Bittner, National Library of Medicine
Leandra Budau, Toronto Metropolitan University
Ioana Buhnila, Center for Data Science in Humanities, Chosun University
Leonardo Campillos-Llanos, Consejo Superior de Investigaciones Cientificas (CSIC, Spanish National Research Council)
Liuliu Chen, The University of Melbourne
Brandon Colelough, National Library of Medicine
Brian Connolly, Cincinnati Children's Hospital Medical Center
Mike Conway, University of Melbourne
An Dao, The University of Tokyo
Oumaima El Khettari, Nantes Université - LS2N
Mohamed Elmofty, Humboldt University of Berlin
Pietro Ferrazzi, University of Padova
Kathleen C. Fraser, University of Ottawa
Natalia Grabar, CNRS STL UMR8163, Université de Lille
Cyril Grouin, LISN
Deepak Gupta, National Library of Medicine, NIH
Thierry Hamon, LISN, Universite Paris-Saclay and Universite Sorbonne Paris Nord
Yikun Han, yikunh2@illinois.edu
Keno Hanken, Independent Researcher
Moustafa Hassan, Doha Institute for Graduate Studies
Sam Henry, Randolph Macon College
Ben Holgate, King's College London
Bahar Ilgen, Robert Koch Institute
Antonio Jimeno Yepes, RMIT University
Indika Kahanda, University of North Florida
Vani Kanjirangat, IDSIA
Sarvnaz Karimi, CSIRO
Nazmul Kazi, University of North Florida

Halil Kilicoglu, University of Illinois Urbana-Champaign
Won Gyu Kim, DIR/NLM/NIH
Ashwin Kirubakaran, Edison Academy Magnet School
Gaurav Kumar, University of California San Diego
Thomas Labbe, Orange / Bcom / CHU Rennes
Andre Lamurias, NOVA School of Science and Technology
Vojtech Lanz, Institute of Formal and Applied Linguistics, Charles University, Czech Republic
Majid Latifi, University of York
Alberto Lavelli, FBK
Robert Leaman, NLM / NIH
Lung-Hao Lee, National Yang Ming Chiao Tung University
Ulf Leser, Humboldt-Universität zu Berlin
Yuan Liang, Queen Mary University of London
Siting Liang, German Research Center for Artificial Intelligence
Livia Lilli, Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Rome, Italy; Catholic University of the Sacred Heart, Rome, Italy
Ying-Jia Lin, Chang Gung University
Jinghui Liu, CSIRO
Fabien Maury, Inserm
Makoto Miwa, Toyota Technological Institute
Rodrigo Morales-Sánchez, Universidad Nacional de Educación a Distancia (UNED)
Hyeonyeong Nam, Department of Artificial Intelligence, Korea University
Claire Nedellec, MaIAGE INRAE
Guenter Neumann, DFKI and Saarland University
Aurélie Névéol, Université Paris Saclay, CNRS, LISN
Brian Ondov, Yale School of Medicine
John E. Ortega, Northeastern University
Olga Pelloni, UiT The Arctic University of Norway
Noon Pokaratsiri Goldstein, DFKI
Juan Prieto, Universidad de Los Andes
François Remy, Parallia AI
Francisco J. Ribadas-Pena, University of Vigo
Fabio Rinaldi, IDSIA USI-SUPSI, Dalle Molle Institute for Artificial Intelligence
Roland Roller, DFKI SLT Lab
Mahule Roy, University of Oxford
Nourah Salem, Computational Bioscience Program, University of Colorado, Anschutz Medical Campus, Aurora, CO, 80045, USA
Vicente Ivan Sanchez Carmona, Ricoh Software Research Center (Beijing) Co., Ltd.
Mustafa Sikder, U.S. Food and Drug Administration
Ujjwal Singh, Max Healthcare Institute Limited
Saurabh Singh, Oracle
Sarvesh Soni, National Library of Medicine
Adam Sutton, King's College London
Mario Sänger, AstraZeneca
Sanya Taneja, Johnson and Johnson Innovative Medicine
Karin Verspoor, RMIT University
Xing David Wang, Computer Science Department, Humboldt-Universität zu Berlin
Nathan M. White, James Cook University; Western Institute for Endangered Language Documentation
Dongfang Xu, Cedars-Sinai Medical Center
Ken Yano, The National Institute of Advanced Industrial Science and Technology

Hyunwoo Yoo, Drexel University
Jingqing Zhang, Pangaea Data
Xiao Yu Cindy Zhang, University of British Columbia
Angelo Ziletti, Bayer AG
Pierre Zweigenbaum, LISN, CNRS, Université Paris-Saclay

Keynote Talk

AI Safety in Healthcare: Ethical and Technical Considerations

Annika Marie Schoene
Bouve College of Health Sciences
2026-07-03 11:00:00 – Room: **Room tbd**

Abstract: The rapid integration of artificial intelligence into healthcare, spanning ambient documentation, virtual nursing, and medical imaging, has outpaced the development of robust oversight mechanisms. While the global AI-in-healthcare market is projected to exceed $868 billion by 2030, incidents of algorithmic bias and unequal care delivery have exposed significant gaps between responsible AI principles and clinical implementation. This talk examines AI safety as a foundational pillar of responsible AI in health contexts, arguing that fairness, transparency, accountability, and human autonomy cannot remain aspirational without corresponding technical and governance infrastructure. I survey existing frameworks and identify a critical missing layer of concrete, operationalizable evaluation methods that bridge ethical principles and technical practice. To address this gap, I present work toward an actionable framework for responsible AI integration in clinical settings, grounded in the AI Ethics Box, a structured taxonomy derived from biomedical ethics. The framework maps ethical domains to specific technical tests, supporting both pre- and post-deployment evaluation. I also describe an ongoing co-design process with clinical and community stakeholders to ensure real-world feasibility. I conclude that AI in healthcare demands genuinely interdisciplinary research to build tools that improve health outcomes without compromising patient safety.

Bio: Annika Marie Schoene, PhD is a computer scientist and researcher in AI safety, working on the evaluation, robustness, and security of large-scale AI systems, including large language models. She develops technical methods and evaluation frameworks to identify and mitigate high-risk behaviors such as jailbreaks, harmful outputs, and unsafe system behavior, with the goal of enabling the safe and trustworthy deployment of AI in public health, health systems, and healthcare settings. She is currently an Assistant Professor in the Department of Public Health and Health Sciences and Technical Lead for the Responsible AI Practice at Northeastern University. Beyond her core research, she works across academic, industry, public-sector and not-for-profit settings to generate evidence that supports non-technical stakeholders and policymakers in making informed decisions that reduce algorithmic harm and inform organizational decision-making. She is also a Visiting Scientist at MaineHealth and the University of Southampton (UK), and a Faculty Fellow at the Institute for Social Justice and Healthy Equity, and serves as a Scientific Expert Advisor at Meta on AI safety. She holds a PhD in Computer Science from the University of Hull (UK). During her doctoral training, she conducted research on machine and deep learning methods for analyzing complex real-world text and interned at IBM Research (UK), continuing this collaboration throughout her PhD and postdoctoral work. She completed her postdoctoral training at the University of Manchester's National Centre for Text Mining (NaCTeM), where she worked on natural language processing in health-related contexts. Prior to her current role, she was a Research Scientist at the Institute for Experiential AI (EAI), working primarily on research involving the development and evaluation of AI methods in applied health contexts.

Table of Contents

Program

Friday, July 3, 2026

08:40 - 08:50 *Opening Remarks*

08:50 - 10:30 *Session 1: Clinical NLP*

Trustworthy NLP for Low-Resource Languages: Agent-Based Uncertainty Modeling for Hebrew Radiology Report Structuring
Hadas Ben Atya, Naama Gavrielov, Zvi Badash, Gili Focht, Ruth Cytter-Kuint, Talar Hagopian, Dan Turner and Moti Freiman

Reliable Automated Triage in Spanish Clinical Notes: A Hybrid Framework for Risk-Aware HIV Suspicion Identification
Rodrigo Morales-Sánchez, Soto Montalvo and Raquel Martínez

Using Synthetic Records to Improve Automated Identification of Seizure Freedom in Clinical Text about People with Epilepsy
Stephen Barlow, Yujian Gan, Joe Davies, Joel Winston, James Teo, Mark Richardson and Ben Holgate

A Multi-Agent Open-Source LLM for Structured Cancer Registry Information Extraction from Pathology and Medical Reports
Abdulrahman Aal Abdulsalam, Adhari Al Zaabi, Riham Jeeballah and Habiba El Keraby

Clinical Evidence and Patient Reviews: A Linked Dataset for Antidepressant Side Effects
Steven Au

Agentic AI Architectures for SOAP Note Generation
Keno Hanken

PromptRad: Knowledge-Enhanced Multi-Label Prompt-Tuning for Low-Resource Radiology Report Labeling
Ying-Jia Lin, Tzu-Chin Lo, Ping-Chien Li, Chi-Tung Cheng, Chien-Hung Liao and Hung-Yu Kao

When Demographic Sensitivity Isn't What It Seems: Baseline-Aware Counterfactual Audits for Clinical NLP
Hyunwoo Yoo

MAX-EVAL-11: A Large Scale Benchmark for Evaluating Large Language Models on Full-Spectrum ICD-11 Medical Coding
Ujjwal Singh, Sarthak Deshwal, Nitish Dube and Arjun Sharma

Reading Between the Lines: Toward Translating Verbose Patient-authored Messages into Clinician-Formulated Questions
Sarvesh Soni, Madeline Bittner and Dina Demner-Fushman

Friday, July 3, 2026 (continued)

10:30 - 11:00 *Coffee Break*

11:00 - 11:30 *Session 2: Invited Talk by Annika Marie Schoene*

11:30 - 12:30 *Session 3: Shared Tasks Overviews*

Overview of the MedGenVidQA 2026 Shared Task on Medical Generative Video Question Answering
Deepak Gupta, Collin Campbell, Pedram Golnari and Dina Demner-Fushman

Overview of the Medical Decision Extraction, Analysis, and Classification Task (MedExACT) of BioNLP 2026
Mohamed Elgaar, Jiali Cheng, Nidhi Vakil, Mehrnaz Sadrolashrafi, Mitra Mohtarami, Adrian Wong, Hadi Amiri and Leo Celi

Overview of the PsyDefDetect Shared Task at BioNLP 2026: Detecting Levels of Psychological Defense Mechanisms in Supportive Conversations
Hongbin Na, Zimu Wang, Zhaoming Chen, Yining Hua, Rena Gao, Kailai Yang, Ling Chen, Wei Wang, Shaoxiong Ji, John Torous and Sophia Ananiadou

Overview of the ClinicalSkillQA 2026 Shared Task on Continuous Perception and Procedural Reasoning in Clinical Skill Assessment
Xiyang Huang, Renxiong Wei, Yihuai Xu, Zhiyuan Chen, Keying Wu, Jiayi Xiang, Buzhou Tang, Yanqing Ye, Jinyu Chen, Cheng Zeng, Min Peng, Qianqian Xie and Sophia Ananiadou

12:30 - 14:00 *Lunch*

14:00 - 15:30 *Session 4: Foundational tasks*

BioConflict: A Benchmark for Evaluating Large Language Models in Biomedical Contradiction Detection and Consensus Synthesis
Ashwin Kirubakaran and Henry Gagnier

Hierarchy-Aware Hyperbolic and Semantic Reranking for Ontology-Based Phenotype Linking
Thomas Labbe, Moussa Baddour, Axel Bonesteve, Paul Rollier, Marie De Tayrac and Olivier Dameron

GRAFT: Gated Retrieval-Augmented Fine-Tuning for Relation Extraction
Yuhang Jiang and Ramakanth Kavuluru

Can NLP Models Detect When One Publication Outweighs Twenty? Predicting Systematic Review Conclusion Changes
Ebrahim Alharbi and Mark Stevenson

Friday, July 3, 2026 (continued)

Divide-Prompt-Refine: a Training-Free, Structure-Aware Framework for Biomedical Abstract Generation
Sylvey Lin, Joseph Menke, Shufan Ming, Dongin Nam, Neil Smalheiser and Halil Kilicoglu

What Do Biomedical NER and Entity Linking Benchmarks Measure? A Corpus-Centric Diagnostic Framework
Robert Leaman, Rezarta Islamaj and Zhiyong Lu

15:30 - 16:00 *Coffee Break*

15:00 - 18:00 *Poster Sessions*

The Divergence Hypothesis: Unmasking Lexical Interference and Label Bias in Mental Health NLP
Moustafa Hassan

Towards Unified Factuality Evaluation for Biomedical QA and Summarization: Aligning Metrics with Clinical Use-Cases
Mahule Roy and Subhas Roy

Analyzing Prompt Design Choices in Biomedical Information Extraction for Low-Resource Languages
Ayesha Khatun, Kadir Bulut Ozler, Steven Bethard and Egoitz Laparra

Agentic Feature Selection via LLM for Epileptic Seizure Detection
Aizierjiang Aiersilan and Xiaodong Qu

Training Biomedical Retrievers From Large-Scale Citation Contexts
Xing David Wang, Duy Le Thanh and Ulf Leser

Gold Label Errors in the SciFact Benchmark: An LLM-Assisted Annotation Audit
Julien Sylvestre

BioRAG: A Systematic Ablation Study of Retrieval Strategies for Biomedical Question Answering
Krushil Bhojani, Mayank Waghmare and Hima Bindu Nandyala

Post Hoc Agentic Refinement for Improving Precision in Multilingual Clinical Text De-identification
Justin Xu, Alistair Johnson, Thomas Lin, David Eyre and Rodolfo Quispe

Do Syntactic Features Help Biomedical Relation Extraction? An Empirical Study of Verb Token and Dependency Graph Augmentation
Mustafa Sikder and Ernest Kwegyir-Afful

Beyond Knowledge Graphs: PubMedBERT Embeddings as a Competitive Standalone Modality for Drug Re-purposing
Rishik Kondadadi and John E. Ortega

CoreELM: An Open-Source Framework for Aligning Large Language Models to Embedding Spaces
Brian Ondov, Chia-Hsuan Chang, Yujia Zhou, Mauro Giuffrè and Hua Xu

Uncertainty-Aware Multi-Label Routing of Clinical Text to Surveillance Pathways
Agathe Zecevic, Sebastian Zeki and Angus Roberts

MedCAT v2: a modular, extensible architecture for clinical named entity recognition and linking under real-world privacy and compute constraints
Mart Ratas, Thomas Searle, Adam Sutton and Richard Dobson

Effects of Adaptive Pretraining in Specialized Domains for Named Entity Recognition
Jack Lynam and Sam Henry

Trade-offs in Medical LLM Adaptation: An Empirical Study in French QA
Ikram Belmadani, Oumaima El Khettari, Carlos Ramisch, Frederic Bechet, Richard Dufour and Benoit Favre

Diagnosing Lower Extremity Arteriovenous Diseases Using Agentic LLMs
Zicen Liao, Yunhao Sun and Matthew Purver

KGRxn-LLM: Knowledge Graph Enhanced Large Language Models for Molecular Reaction Reasoning
Weichen Liu, Qiyao Xue, Yuyang Wu, Olexandr Isayev and Natasa Miskov-Zivanov

Treating Decoder-Only LLMs as Encoders: A Simple and Effective Fine-tuning Approach for Named Entity Recognition
Ken Yano and Hiroya Takamura

A Multi-View Framework for Cross-Domain Nutrition Misinformation Detection in Social Media
Vishwaa Shah, Indika Kahanda, Andrea Arikawa, Asal Abbaszadeh and Richard Loftis

Ontological Validation of Biomedical Topic Models: SNOMED CT Hierarchy Distance as an Automated Evaluation Metric
Ilan Rubinfeld, Sami Zaidi, Milosh Djuric, Loay Kabbani, Mouhammad Halabi and Alex Shepard

Systematic Evaluation of the Quality of Synthetic Clinical Notes Rephrased by LLMs at Million-Note Scale
Jinghui Liu, Sarvesh Soni and Anthony Nguyen

Bridging the Version Gap: Multi-version Training Improves ICD Code Prediction, Especially for Rare Codes
Jinghui Liu and Anthony Nguyen

EmCellLLM: Human Peri-Implantation Embryonic Cell Annotation Based on Large Language Models
Xiaorui Guo, Zhiwei Liu, Qianqian Xie and Sophia Ananiadou

Randomized Controlled Trials as the Gold-Standard for Evaluating LLMs: A Primer for Biomedical NLP Researchers
Vicente Ivan Sanchez Carmona, Shanshan Jiang and Bin Dong

Citation-Aware Continual Pre-Training for Biomedical Language Models
Masaki Asada, Tomoki Tsujimura, Tatsuya Ishigaki, Shusaku Egami, Ken Fukuda and Hiroya Takamura

TrackList: Tracing Back Query Linguistic Diversity for Head and Tail Medical Knowledge in Open Large Language Models
Ioana Buhnila, Aman Sinha and Mathieu Constant

Discharge Instructions are not One Task: Grounding Differences Between Surgical and Non-Surgical Admissions
Mayank Jobanputra, Justin Xu, Samarth Oza, Hulma Naseer, Yifan Wang, Blerta Veseli, Chandralekha Kona, Haochen Cui, David Eyre and Vera Demberg

PrionNER: A Named Entity Recognition Dataset for Prion Disease Biomedical Literature
An Dao, Nhan Ly, Thao Tran, Yuji Matsumoto and Akiko Aizawa

Evaluation of Multilingual Text Simplification for the Mental Health Domain: Exploring Small Language Models
Olga Pelloni, Sandra Just and Lars Bongo

BioTopicXplor: A Web Tool for Interactive Exploration of PubMed Literature through Reproducible Topics.
Lana Yeganova, Donald Comeau, Won Kim, Natalie Xie, Shubo Tian, W John Wilbur and Zhiyong Lu

Investigating Stigmatizing Language in Clinical Documentation with Open-Source Large Language Models
Rajashree Dahal, Pardis Hosseinpour, Pranithi Kamishetty, Satwik Pamulaparthy, Saeid Tizpaz-Niari and Natalie Parde

Learning to Combine AI Annotations for Improved Biomedical Relevance Labeling
Won Gyu Kim, Lana Yeganova, Shubo Tian, Donald Comeau, W John Wilbur and Zhiyong Lu

When Does Retrieval Beat Direct LLM Diagnosis in Rare Disease? An Empirical Study of Ontology Coverage
Mohamed Elmofty and Ulf Leser

BioCoref: Benchmarking Biomedical Coreference Resolution with LLMs
Nourah Salem, Elizabeth White, Michael Bada and Lawrence Hunter

Tokenization Granularity and Medical Term Representations in Language Models
Vojtech Lanz and Pavel Pecina

CAP: A Source-Grounded Proposition Scaffold for Faithful Clinical Dialogue-to-Note Generation
Hyunkyung Lee, Jisoo Jung, Jeonguk Lee, Jaehyo Yoo, Wooseok Han, Minkyu Kim and Gibaeg Kim

Segmentation Matters: Exploring LLM-Based Strategies for Temporal Clinical Event Identification in Oncology Reports
Cristiano Bellucci, Francesco Madeddu, Chiara Iacomini, Carlotta Masciocchi, Stefano Patarnello, Massimo Bernaschi, Mario Santoro and Livia Lilli

Operation-Mechanism Alignment for Reliable Clinical Reasoning over Electronic Health Records
Guanyu Tao, Siyao Wang, Yong Xue, Ashwani Tanwar, Yuting Ji, Kai Sun, Monica Mok, Marzana Chowdhury, Deepa Gupta, Ashok Gupta, Jingqing Zhang, Vibhor Gupta and Yike Guo

MeSHClass-ES and AnatEM-ES: Open Resources for Spanish Biomedical NLP
Santiago Martinez Novoa, Lina Gomez Mesa, Juan Prieto and Ruben Manrique

When Evidence Conflicts: Uncertainty and Order Effects in Retrieval-Augmented Biomedical Question Answering
Yikun Han, Mengfei Lan and Halil Kilicoglu

A Comparative Analysis of In-Context Learning and Fine-Tuning for Biomedical Information Retrieval and Sentence Extraction Using Research Domain Criteria
Athlene Jones, Khanh Lieu and Indika Kahanda

A Deterministic Multi-Stage Retrieval Pipeline for Longitudinal EHR Question Answering
Shubham Agarwal, Thomas Searle, Richard Dobson, Ninoslav Majkic and Niko Moller-Grell

Interpretable ICD Code Classification with Faithful Sentence Extraction
Yichen Wang, Lian Hong, Masato Mizogaki, Shunnosuke Umeda, Toshimune Kenmotsu, Akihiro Tamura and Daniel Andrade

Evaluating LLM-as-a-Judge for Medical Term Simplification
Ioana Buhnila, Aman Sinha, Rohit Agarwal, Dilip K. Prasad and Mathieu Constant

FACT: Functional Group Alignment and Consistency in Token Space for Structure-aware Molecular Representation Learning
Hyeonyeong Nam, Woojae Choi, Deok-Joong Lee, Young-Han Son, Sangwoon Lee, Bogyeong Kang, Eunjung Jo and Tae-Eui Kam

Small LLMs for Biomedical Claim Verification: Cost-Effective Fine-Tuning, Structural Dataset Shortcuts, and Cross-Domain Generalization
Gaurav Kumar

Diagnosable ColBERT: Debugging Late-Interaction Retrieval Models Using a Learned Latent Space as Reference
François Remy

Developing Literature Annotation Guidelines for Representing Normal Physiology in Biolink-Compatible Knowledge Graphs
Madeline Bittner, Willie Rogers, Dina Demner-Fushman, Richard Scheuermann and Matthew Diller

CENT: Context Engineering Framework for Normalization of Clinical Trial Procedures
Sanya Taneja, Ziqing Ji, Hans Verstraete and Ali Samadani

VERICITE: Evaluating Sentence-Level Citation Faithfulness in Retrieval-Augmented Medical Question Answering
Yixian Ma, Bohao Chu and Norbert Fuhr

AAbAAC: An Annotated Corpus for Autoimmunity Information Extraction
Fabien Maury, Solène Grosdidier, Maud De Dieuleveult and Adrien Coulet

Towards Grounded Hallucination Definitions for Biomedical Question Answering with Reproducible Examples from ClinIQLink
Brandon Colelough, Davis Bartels, Madeline Bittner and Dina Demner-Fushman

VaxScope: Document-Level Structured Evidence Extraction from Immunization Systematic Reviews
Bahar Ilgen, Ebenezer Awotoro and Georges Hattab

Medical Context Variation: A source of impairment for Event classification
Aman Sinha, Marianne Clausel, Mathieu Constant and Xavier Coubez

KALIMBA: Knowledge-Assisted Literature Mining for Biological Interaction Analysis
Niloofar Arazkhani, Maciej Kotecki, Brent Cochran and Natasa Miskov-Zivanov

When Retrieval Doesn't Help: A Large-Scale Study of Biomedical RAG
Erfan Nourbakhsh, Rocky Slavin, Ke Yang and Anthony Rios

CrossDDI: Cross-Source Evidence-Grounded Drug-Drug Interaction Verification
Bohao Chu and Norbert Fuhr

SCoPE: Planning for Hybrid Querying over Clinical Trial Data
Suparno Chowdhury, Manan Choudhury, Tejas Anvekar, Muhammed Khan, Kaneez Khakwani, Mohamad Sonbol, Irbaz Riaz and Vivek Gupta

Expert-Guided Schema-Based Structured Extraction from CONSORT Diagrams Using Vision-Language Models
Damian Stachura, Bartosz Przechera, Monika Opa?ek, Ewelina Sadowska, Ewa Borowiack and Artur Nowak

From Rules to Predictions: Federated Tabular Learning with LLM Reasoning
Afsaneh Mahanipour and Hana Khamfroush

MedBench: Deliberative Evaluation of Medical Language Models
Pratik Jalan, Mukul Joshi, Akhilesh Magotra and Kshitij Jadhav

Fast, Accurate, and Local Conversion of MIMIC-IV to OMOP with DBT
Adam Sutton, Niko Moller-Grell, Thomas Searle and Richard Dobson

Exploring Novel Drug Research Area using Large Language Models Based on Research Trends in Biomedical Literature
Afnan Afnan, Michael Van Supranes, Tomohiro Nishiyama, Shoko Wakamiya and Eiji Aramaki

FHexchange: Resources for Family Health History Extraction and Normalization From Consumer Dialog Sources
Michelle Nguyen, Nidhi Soley, Ayah Zirikly, João Sedoc and Casey Taylor

Forgotten Words: Benchmarking NeoBERT for Dementia Detection in Low-Resource Conversational Filipino and English Speech
Rez Samantha Floresca, Edric Castel Hao, Hannah Grachiella Buñales, Chelsea Dominique Temprosa, Georgianna Reyes and Kervin Gabriel Chua

IndicMedDialog: A Parallel Multi-Turn Medical Dialogue Dataset for Accessible Healthcare in Indic Languages
Shubham Nigam, Suparnojit Sarkar and Piyush Patel

Towards a Radiologist Imitation Framework for 3D CT Diagnosis with Multimodal LLMs
Kaidi Zhang, Zhiyuan Yan, Gao Cheng and Zhenyang Cai

Probing and Steering Uncertainty in Biomedical Language Models: Representational Structure and Behavioral Limits
Debmalya Pal

Relations of Linguistic Features and Medical Text Preferences are Nontrivial
Davis Bartels, Brandon Colelough and Dina Demner-Fushman

17:50 - 18:00 *Closing Remarks*