ACL 2025

BioNLP 2025 and Shared Tasks

Proceedings of the 24th Workshop on Biomedical Language Processing

August 1, 2025

The ACL organizers gratefully acknowledge the support from the following sponsors.

In cooperation with

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 979-8-89176-275-6

BioNLP 2025: new solutions to perennial and emerging problems

Dina Demner-Fushman, Sophia Ananiadou, Makoto Miwa and Jun-ichi Tsujii

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 2025, 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 drug discovery and gene editing to veterinary and 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 2025 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-professional medical sub-languages, among others. BioNLP remains the flagship and the generalist in biomedical language processing, accepting all noteworthy work independently of the tasks and languages studied. The quality of submissions continues to impress the program committee and the organizers.

BioNLP 2025 received 61 submissions, of which eight were accepted for oral presentation and 22 as poster presentations. The selected works span foundational research, biomedical language processing, clinical applications, and generation of new datasets and benchmarks.

Four Shared Tasks were collocated with BioNLP 2025:

SMAFIRA: annotating the literature for finding methods alternative to animal experiments.

ClinIQLink 2025: LLM Lie Detector Test: evaluating the effectiveness of generative models in producing factually accurate information, using a benchmark dataset specifically curated to align with the knowledge level of a General Practitioner (GP) .

ArchEHR-QA 2025: Grounded Electronic Health Record Question Answering: automatically generating answers to patients' health-related questions that are grounded in the evidence from patients' clinical notes.

BioLaySumm 2025: Now, in its third edition, this year's BioLaySumm, introduces a new task: radiology report generation in layman's terms, extending the shared task to a new multimodal domain.

The overviews of the tasks and short presentations of the best performing approaches are included in the workshop program. The participants in all Shared Tasks present their work in a dedicated poster session.

The keynote by Wojciech Kusa is titled: Incorporating Changes in Review Outcomes in the Evaluation of Systematic Review Automation.
Current evaluations of automation methods in systematic literature reviews often treat all included studies as equally important, ignoring their varying influence on review outcomes. This can misrepresent the effectiveness of search strategies, as not all relevant studies contribute equally to the conclusions of the review. To address this limitation, we propose a new evaluation framework that incorporates the differential impact of individual studies on review outcomes. Using data from the CLEF 2019 TAR task, we applied this framework to assess 74 automation models, leveraging meta-analysis effect estimates to weigh the influence of each study. Compared to conventional binary relevance metrics, our approach provided a more nuanced assessment, emphasizing the importance of retrieving high-impact studies. Results showed significant differences in model rankings, underscoring the value of outcome-based evaluation.

This framework offers researchers a more precise method for evaluating systematic review automation tools, ultimately supporting higher-quality evidence synthesis and better-informed clinical decisions.

Wojciech is a Senior Researcher at the NASK National Research Institute in Poland, where he leads the Linguistic Engineering and Text Analysis Department. He holds a PhD in NLP from TU Wien, with a focus on applying and evaluating neural methods for domain-specific data. His research interests include the safety and evaluation of large language models, clinical and biomedical NLP, and AI-driven scientific discovery. Wojciech was a Marie Skłodowska-Curie Fellow in the EU Horizon 2020 project DoSSIER, specialising in biomedical information retrieval and NLP. He has industry experience from roles at Samsung and Allegro, and has completed research internships at Sony, UNINOVA, and the Polish Academy of Sciences.

We are pleased to announce that the Chen Institute is co-organizing the BioNLP 2025 Workshop. Founded in 2016 by Tianqiao Chen and Chrissy Luo, the Chen Institute is driven by a bold vision to improve the human experience by understanding how our brains perceive, learn, and interact with the world. Their global platform includes the Tianqiao and Chrissy Chen Institute for Neuroscience at Caltech, the Tianqiao Chen Institute for Translational Research in Shanghai, the Chen Frontier Lab for Applied Neurotechnology, and the Chen Frontier Lab for AI and Mental Health. The Chen Scholars program supports early- to mid-career scientists, and the recently launched Chen Institute and Science Prize for AI Accelerated Research highlights their deep commitment to innovation. At this year's BioNLP Workshop, the Chen Institute is interested in exploring how artificial intelligence can accelerate the pace of scientific discovery. We believe there are vast, untapped opportunities to make groundbreaking advances by leveraging the power of AI. The hope is that this meeting will serve as the beginning of an ongoing dialogue—focused on new developments, transformative successes, and emerging thinking at the intersection of AI and science. Through this collaboration, the Chen Institute aims to identify and support promising approaches with the potential to meaningfully change the world.

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 fairly short review period. The quality of submitted work continues to grow, and the organizers are truly grateful to the members of our amazing 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.

Organizing Committee

Chair

Dina Demner-Fushman, National Library of Medicine, USA

Organizers

Sophia Ananiadou, National Centre for Text Mining and University of Manchester, UK
Makoto Miwa, Toyota Technological Institute, Japan
Junichi Tsujii, National Institute of Advanced Industrial Science and Technology, Japan

Program Committee

Chairs

Dina Demner-Fushman, National Library of Medicine
Sophia Ananiadou, University of Manchester
Makoto Miwa, Toyota Technological Institute
Junichi Tsujii, National Institute of Advanced Industrial Science and Technology, Japan

Program Committee

Natalie Alexander, The University of Cape Town
Daniel Andrade, Hiroshima University
Emilia Apostolova, Language.ai
Eiji Aramaki, NAIST, Japan
Tanmay Basu, Indian Institute of Science Education and Research Bhopal
Leandra Budau, Toronto Metropolitan University
Leonardo Campillos-Llanos, Consejo Superior de Investigaciones Cientificas (Spanish National Research Council)
Yingjian Chen, Henan University
Liuliu Chen, The University of Melbourne
Brian Connolly, Cincinnati Children's Hospital Medical Center
Mike Conway, University of Melbourne
An Dao, The University of Tokyo
Berry De Bruijn, National Research Council Canada
Jean-Benoit Delbrouck, Stanford University
Simona Doneva, University of Zurich
Pietro Ferrazzi, University of Padova
Kathleen C. Fraser, National Research Council Canada
Tomas Goldsack, University of Sheffield
Natalia Grabar, CNRS STL UMR8163, Université de Lille
Cyril Grouin, LIMSI-CNRS
Tudor Groza, Pryzm Health Pty Ltd
Yingjun Guan, iSchool, University of Illinois at Urbana-Champaign
Deepak Gupta, National Library of Medicine, NIH
Thierry Hamon, LISN, Universite Paris-Saclay
Universite Sorbonne Paris Nord
William Hogan, UCSD
Ben Holgate, King's College London
Brian Hur, University of Washington
Antonio Jimeno Yepes, Unstructured Technologies
Hidetaka Kamigaito, Nara Institute of Science and Technology
Vani Kanjirangat, IDSIA
Sarvnaz Karimi, CSIRO
Nazmul Kazi, University of North Florida
Siun Kim, Seoul Natoinal University Hospital
Gaurav Kumar, University of California San Diego
Andre Lamurias, NOVA School of Science and Technology
Majid Latifi, University of York
Alberto Lavelli, FBK

Robert Leaman, National Center for Biotechnology Information
L u n g - H a o 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
Abdine Maiga, University College London
Claire Nedellec, INRAE
Guenter Neumann, DFKI
Saarland University
Mariana Neves, German Federal Institute for Risk Assessment
Andrei Niculae, National University of Science and Technology Politehnica Bucharest
Aurélie Névéol, Université Paris Saclay, CNRS, LISN
Brian Ondov, Yale School of Medicine
Noon Pokaratsiri Goldstein, DFKI
François Remy, Ghent University
Francisco J. Ribadas-Pena, University of Vigo
Fabio Rinaldi, IDSIA, Swiss AI Institute
Roland Roller, DFKI SLT Lab
Mourad Sarrouti, CLARA Analytics
Efstathia Soufleri, Athena RC
Peng Su, University of Delaware
Madhumita Sushil, University of Antwerp
Mario Sänger, Humboldt-Universität zu Berlin
Andrew Taylor, Yale University
Karin Verspoor, RMIT University
Davy Weissenbacher, Cedars-Sinai Medical Center
Nathan M. White, James Cook University; Western Institute for Endangered Language Documentation
Dongfang Xu, Cedars-Sinai Medical Center
Shweta Yadav, University of Illinois at Chicago
Ken Yano, The National Institute of Advanced Industrial Science and Technology
Hyunwoo Yoo, Drexel University
Xiao Yu Cindy Zhang, University of British Columbia
Xinyue Zhang, King's College London
Kai Zhang, Worcester Polytechnic Institute
Jingqing Zhang, Pangaea Data
Angelo Ziletti, Bayer AG
Ayah Zirikly, Johns Hopkins University
Pierre Zweigenbaum, LISN, CNRS, Université Paris-Saclay

Secondary Reviewers

Joseph Akinyemi, University of York
Robert Bossy, National Research Institute for Agriculture, Food and Environment (INRAE)
Marco Naguib, Interdisciplinary Laboratory on Numerical Sciences (LISN)

Keynote Talk

Incorporating Changes in Review Outcomes in the Evaluation of Systematic Review Automation

Wojciech Kusa
NASK National Research Institute, Poland
2025-08-01 12:00:00 – Room: **Room 2.15**

Abstract: Current evaluations of automation methods in systematic literature reviews often treat all included studies as equally important, ignoring their varying influence on review outcomes. This can misrepresent the effectiveness of search strategies, as not all relevant studies contribute equally to the conclusions of the review. To address this limitation, we propose a new evaluation framework that incorporates the differential impact of individual studies on review outcomes. Using data from the CLEF 2019 TAR task, we applied this framework to assess 74 automation models, leveraging meta-analysis effect estimates to weigh the influence of each study. Compared to conventional binary relevance metrics, our approach provided a more nuanced assessment, emphasizing the importance of retrieving high-impact studies. Results showed significant differences in model rankings, underscoring the value of outcome-based evaluation. This framework offers researchers a more precise method for evaluating systematic review automation tools, ultimately supporting higher-quality evidence synthesis and better-informed clinical decisions.

Bio: Wojciech is a Senior Researcher at the NASK National Research Institute in Poland, where he leads the Linguistic Engineering and Text Analysis Department. He holds a PhD in NLP from TU Wien, with a focus on applying and evaluating neural methods for domain-specific data. His research interests include the safety and evaluation of large language models, clinical and biomedical NLP, and AI-driven scientific discovery. Wojciech was a Marie Skłodowska-Curie Fellow in the EU Horizon 2020 project DoSSIER, specialising in biomedical information retrieval and NLP. He has industry experience from roles at Samsung and Allegro, and has completed research internships at Sony, UNINOVA, and the Polish Academy of Sciences.

Table of Contents

Program

Friday, August 1, 2025

08:40 - 08:50 *Opening Remarks*

08:50 - 10:30 *Session 1: Foundational tasks*

Accelerating Cross-Encoders in Biomedical Entity Linking
Javier Sanz-Cruzado and Jake Lever

Beyond Citations: Integrating Finding-Based Relations for Improved Biomedical Article Representations
Yuan Liang, Massimo Poesio and Roonak Rezvani

MedSummRAG: Domain-Specific Retrieval for Medical Summarization
Guanting Luo and Yuki Arase

Advancing Biomedical Claim Verification by Using Large Language Models with Better Structured Prompting Strategies
Siting Liang and Daniel Sonntag

Questioning Our Questions: How Well Do Medical QA Benchmarks Evaluate Clinical Capabilities of Language Models?
Siun Kim and Hyung-Jin Yoon

10:30 - 11:00 *Coffee Break*

11:00 - 12:30 *Session 2: Clinical NLP*

A Retrieval-Based Approach to Medical Procedure Matching in Romanian
Andrei Niculae, Adrian Cosma and Emilian Radoi

Error Detection in Medical Note through Multi Agent Debate
Abdine Maiga, Anoop Shah and Emine Yilmaz

Converting Annotated Clinical Cases into Structured Case Report Forms
Pietro Ferrazzi, Alberto Lavelli and Bernardo Magnini

Friday, August 1, 2025 (continued)

12:00 - 12:30 *Session 3: Invited Talk by Wojciech Kusa*

12:30 - 14:00 *Lunch*

14:00 - 15:30 *Session 4: Shared Tasks*

14:00 - 14:15 *BioLaySum*

Overview of the BioLaySumm 2025 Shared Task on Lay Summarization of Biomedical Research Articles and Radiology Reports
Chenghao Xiao, Kun Zhao, Xiao Wang, Siwei Wu, Sixing Yan, Tomas Goldsack, Sophia Ananiadou, Noura Al Moubayed, Liang Zhan, William K. Cheung and Chenghua Lin

14:15 - 14:25 *BioLaySum Poster Boasters*

14:25 - 14:40 *SMAFIRA*

SMAFIRA Shared Task at the BioNLP'2025 Workshop: Assessing the Similarity of the Research Goal
Mariana Neves, Iva Sovadinova, Susanne Fieberg, Celine Heinl, Diana Rubel, Gilbert Schönfelder and Bettina Bert

14:40 - 14:55 *CliniIQLink*

Overview of the ClinIQLink 2025 Shared Task on Medical Question-Answering
Brandon Colelough, Davis Bartels and Dina Demner-Fushman

14:55 - 15:00 *CliniIQLink Poster Boasters*

15:00 - 15:15 *ArchEHR-QA*

Overview of the ArchEHR-QA 2025 Shared Task on Grounded Question Answering from Electronic Health Records
Sarvesh Soni, Soumya Gayen and Dina Demner-Fushman

15:15 - 15:25 *ArchEHR-QA Poster Boasters*

Friday, August 1, 2025 (continued)

15:30 - 16:00 *Coffee Break*

16:00 - 18:00 *Poster Sessions (see Shared Task posters in Volume 2)*

Improving Barrett's Oesophagus Surveillance Scheduling with Large Language Models: A Structured Extraction Approach
Xinyue Zhang, Agathe Zecevic, Sebastian Zeki and Angus Roberts

Effective Multi-Task Learning for Biomedical Named Entity Recognition
João Ruano, Gonçalo Correia, Leonor Barreiros and Afonso Mendes

PetEVAL: A veterinary free text electronic health records benchmark
Sean Farrell, Alan Radford, Noura Al Moubayed and Peter-John Noble

Can Large Language Models Classify and Generate Antimicrobial Resistance Genes?
Hyunwoo Yoo, Haebin Shin and Gail Rosen

Overcoming Data Scarcity in Named Entity Recognition: Synthetic Data Generation with Large Language Models
An Dao, Hiroki Teranishi, Yuji Matsumoto, Florian Boudin and Akiko Aizawa

Fine-tuning LLMs to Extract Epilepsy Seizure Frequency Data from Health Records
Ben Holgate, Joe Davies, Shichao Fang, Joel Winston, James Teo and Mark Richardson

Transformer-Based Medical Statement Classification in Doctor-Patient Dialogues
Farnod Bahrololloomi, Johannes Luderschmidt and Biying Fu

PreClinIE: An Annotated Corpus for Information Extraction in Preclinical Studies
Simona Doneva, Hanna Hubarava, Pia Härvelid, Wolfgang Zürrer, Julia Bugajska, Bernard Hild, David Brüschweiler, Gerold Schneider, Tilia Ellendorff and Benjamin Ineichen

QoLAS: A Reddit Corpus of Health-Related Quality of Life Aspects of Mental Disorders
Lynn Greschner, Amelie Wührl and Roman Klinger

Gender-Neutral Large Language Models for Medical Applications: Reducing Bias in PubMed Abstracts
Elizabeth Schaefer and Kirk Roberts

LLMs as Medical Safety Judges: Evaluating Alignment with Human Annotation in Patient-Facing QA
Yella Diekmann, Chase Fensore, Rodrigo Carrillo-Larco, Eduard Castejon Rosales, Sakshi Shiromani, Rima Pai, Megha Shah and Joyce Ho

AdaBioBERT: Adaptive Token Sequence Learning for Biomedical Named Entity Recognition
Sumit Kumar and Tanmay Basu

Enhancing Stress Detection on Social Media Through Multi-Modal Fusion of Text and Synthesized Visuals
Efstathia Soufleri and Sophia Ananiadou

MuCoS: Efficient Drug–Target Discovery via Multi-Context-Aware Sampling in Knowledge Graphs
Haji Gul, Abdul Naim and Ajaz Bhat

Enhancing Antimicrobial Drug Resistance Classification by Integrating Sequence-Based and Text-Based Representations
Hyunwoo Yoo, Bahrad Sokhansanj and James Brown

Effect of Multilingual and Domain-adapted Continual Pre-training on Few-shot Promptability
Ken Yano and Makoto Miwa

Understanding the Impact of Confidence in Retrieval Augmented Generation: A Case Study in the Medical Domain
Shintaro Ozaki, Yuta Kato, Siyuan Feng, Masayo Tomita, Kazuki Hayashi, Wataru Hashimoto, Ryoma Obara, Masafumi Oyamada, Katsuhiko Hayashi, Hidetaka Kamigaito and Taro Watanabe

Prompting Large Language Models for Italian Clinical Reports: A Benchmark Study
Livia Lilli, Carlotta Masciocchi, Antonio Marchetti, Giovanni Arcuri and Stefano Patarnello

CaseReportCollective: A Large-Scale LLM-Extracted Dataset for Structured Medical Case Reports
Xiao Yu Cindy Zhang, Melissa Fong, Wyeth Wasserman and Jian Zhu

RadQA-DPO: A Radiology Question Answering System with Encoder-Decoder Models Enhanced by Direct Preference Optimization
Md Sultan Al Nahian and Ramakanth Kavuluru

Benchmarking zero-shot biomedical relation triplet extraction across language model architectures
Frederik Gade, Ole Lund and Marie Lisandra Mendoza

Friday, August 1, 2025 (continued)

Virtual CRISPR: Can LLMs Predict CRISPR Screen Results?
Steven Song, Abdalla Abdrabou, Asmita Dabholkar, Kastan Day, Pavan Dharmoju, Jason Perera, Volodymyr Kindratenko and Aly Khan

17:50 - 18:00 *Closing Remarks*