ACL 2025

The 24th BioNLP Workshop and Shared Tasks

Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks)

August 1, 2025

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 979-8-89176-276-3

Introduction

Sarvesh Soni, Dina Demner-Fushman

Four shared tasks were organized as part of the BioNLP 2025 Workshop. These tasks are designed to advance the state of the art in biomedical natural language processing by providing a platform to foster innovative solutions to challenging problems in the field. Specifically, the tasks addressed: (1) validating the similarity of research goals between scientific articles (SMAFIRA), (2) evaluating the factual accuracy of generative models (ClinIQLink), (3) generating evidence-grounded answers from electronic health records (ArchEHR-QA), and (4) producing lay summaries of biomedical and radiology texts (BioLaySumm). Collectively, these shared tasks foster the development and benchmarking of innovative methods for information retrieval, knowledge assessment, question answering, and summarization in biomedicine.

A total of 35 papers were submitted across the four tasks, with participants detailing a range of novel approaches and systems. As is typical with shared task tracks, the majority of submissions were accepted, resulting in 34 papers being included in the proceedings.

We provide a brief overview of each shared task below:

1. SMAFIRA

The SMAFIRA task addresses the challenge of assessing the similarity of research goals between biomedical articles, a crucial step in identifying alternatives to animal experiments. Participants were provided with a set of reference scientific articles from PubMed, each representing an animal study within a specific disease domain, along with a set of twenty candidate articles retrieved from PubMed. The task involved validating the candidate articles either automatically, with systems of the participants' choice, or manually, using the SMAFIRA web tool. Validation involves comparing titles and abstracts to assess the degree of similarity in research goals using a three-point scale: similar, uncertain, or not similar. This collaborative annotation effort aims to produce a high-quality dataset for benchmarking automated methods and supporting the broader adoption of non-animal research approaches.

2. ClinIQLink

The ClinIQLink task aims to evaluate the ability of generative language models to produce factually accurate medical information. Participants were tasked to submit models that can answer a diverse set of clinically relevant questions, spanning fundamental concepts in procedures, conditions, diagnostics, and pharmacology, at the level expected of a General Practitioner. Using a novel, expert-curated dataset of atomic question-answer pairs, the task assesses both closed- and open-ended responses, employing scoring metrics to measure knowledge retrieval and penalize factual inaccuracies. Beyond benchmarking the current capabilities of generative models, the task provides insights into the origins and types of hallucinations exhibited by state-of-the-art language models in medical contexts.

3. ArchEHR-QA

The ArchEHR-QA task targets the challenge of generating accurate, evidence-grounded answers to patients' health-related questions using information from electronic health records (EHRs). Participants were provided with realistic patient questions, clinician-interpreted version of the questions, and corresponding clinical note excerpts. The objective is to develop systems that generate concise, professional responses that are explicitly supported using citations to relevant sentences from the clinical notes. System outputs are evaluated using two main criteria: factuality, which measures the precision and recall of cited evidence against manually annotated ground truth, and relevance, which assesses the quality and

appropriateness of the generated answer text compared to ground truth. This task aims to advance research on supporting clinicians in efficiently addressing patient inquiries, while ensuring that responses remain accurate, contextually appropriate, and grounded in real clinical evidence.

4. BioLaySumm

The BioLaySumm task focuses on the challenge of making complex scientific information accessible to non-expert audiences. The task comprises two main subtasks. In one, participants were given biomedical scientific articles and tasked to develop systems to produce readable, informative summaries suitable for the general public, with an additional subtask requiring the integration of external knowledge to fill information gaps for lay readers. In the second, participants were tasked to translate professional radiology reports into layman's terms, with an additional multi-modal subtask involving the generation of lay summaries directly from medical images. This task aims to benchmark current approaches and foster the development of systems that support more inclusive and effective biomedical communication.

We remain deeply grateful to all shared task participants, to the authors who submitted papers, and to the reviewers (listed under Program Committee) who provided thorough and thoughtful reviews for the submissions, often under tight timelines. The quality of work submitted continues to rise, and we are indebted to the outstanding members of our Organizing Committee, whose careful assessments have been instrumental in identifying research ready for presentation and in advising authors where further experiments and analyses could strengthen their contributions.

As in previous years, we look forward to a productive workshop and to the new collaborations and research directions it will inspire. We are confident that these efforts will help our community continue to advance public health and well-being, as well as contribute meaningfully to both basic and clinical research.

Organizing Committee

SMAFIRA Shared Task

Mariana Neves, German Federal Institute for Risk Assessment - BfR, Germany

ClinIQLink Shared Task

Brandon Colelough, National Library of Medicine, USA
Dina Demner-Fushman, National Library of Medicine, USA
Davis Bartels, National Library of Medicine, USA

ArchEHR-QA Shared Task

Sarvesh Soni, National Library of Medicine, USA
Dina Demner-Fushman, National Library of Medicine, USA

BioLaySumm Shared Task

Kun Zhao, University of Pittsburgh
Liang Zhan, University of Pittsburgh
Chenghao Xiao, University of Durham
Noura Al Moubayed, University of Durham
Kejing Yin, Hong Kong Baptist University
Sixing Yan, Hong Kong Baptist University
Zijian Lei, Hong Kong Baptist University
William Cheung, Hong Kong Baptist University
Qianqian Xie, Yale University
Zheheng Luo, University of Manchester
Sophia Ananiadou, University of Manchester
Tomas Goldsack, University of Sheffield
Siwei Wu, University of Manchester
Xiao Wang, University of Manchester
Chenghua Lin, University of Manchester

Program Committee

Reviewers

Rebecca Allen, Mount St. Joseph Univ. Center for IT Engagement
Mohammad Arvan, University of Illinois at Chicago
Sai Prasanna Teja Reddy Bogireddy, University of Chicago
Surabhi Datta, IMO Health, USA
Viswanatha Reddy Gajjala, Amazon
Shohreh Haddadan, Moffitt Cancer Center
Ming Huang, UTHealth
Sy Hwang, University of Pennsylvania
Tuan Dung Le, University of South Florida
Jinghui Liu, CSIRO
Adam Remaki, Sorbonne Univserity
Suveyda Yeniterzi, GenAIus Technologies

Table of Contents