

ACL 2025

The 24th BioNLP Workshop and Shared Tasks

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

August 1, 2025

©2025 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 979-8-89176-276-3

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 (BioLay-Summ). 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 Univerity

Suveyda Yeniterzi, GenAIus Technologies

Table of Contents

<i>ArgHiTZ at ArchEHR-QA 2025: A Two-Step Divide and Conquer Approach to Patient Question Answering for Top Factuality</i>	
Adrian Cuadron Cortes, Aimar Sagasti, Maitane Urruela, Iker De La Iglesia, Ane García Domingo-aldama, Aitziber Atutxa Salazar, Josu Goikoetxea and Ander Barrena	1
<i>UNIBUC-SD at ArchEHR-QA 2025: Prompting Our Way to Clinical QA with Multi-Model Ensembling</i>	
Dragos Ghinea and Ștefania Rîncu	11
<i>Loyola at ArchEHR-QA 2025: Exploring Unsupervised Attribution of Generated Text: Attention and Clustering-Based Methods</i>	
Rohan Sethi, Timothy Miller, Majid Afshar and Dmitriy Dligach	22
<i>CUNI-a at ArchEHR-QA 2025: Do we need Giant LLMs for Clinical QA?</i>	
Vojtech Lanz and Pavel Pecina	27
<i>WisPerMed at ArchEHR-QA 2025: A Modular, Relevance-First Approach for Grounded Question Answering on Electronic Health Records</i>	
Jan-Henning Büns, Hendrik Damm, Tabea Pakull, Felix Nensa and Elisabeth Livingstone	41
<i>heiDS at ArchEHR-QA 2025: From Fixed-k to Query-dependent-k for Retrieval Augmented Generation</i>	
Ashish Chouhan and Michael Gertz	50
<i>UniBuc-SB at ArchEHR-QA 2025: A Resource-Constrained Pipeline for Relevance Classification and Grounded Answer Synthesis</i>	
Sebastian Balmus, Dura Bogdan and Ana Sabina Uban	62
<i>KR Labs at ArchEHR-QA 2025: A Verbatim Approach for Evidence-Based Question Answering</i>	
Adam Kovacs, Paul Schmitt and Gabor Recski	69
<i>LAILab at ArchEHR-QA 2025: Test-time scaling for evidence selection in grounded question answering from electronic health records</i>	
Tuan Dung Le, Thanh Duong, Shohreh Haddadan, Behzad Jazayeri, Brandon Manley and Thanh Thieu	75
<i>UTSA-NLP at ArchEHR-QA 2025: Improving EHR Question Answering via Self-Consistency Prompting</i>	
Sara Shields-Menard, Zach Reimers, Joshua Gardner, David Perry and Anthony Rios	81
<i>UTSamuel at ArchEHR-QA 2025: A Clinical Question Answering System for Responding to Patient Portal Messages Using Generative AI</i>	
Samuel Reason, Liwei Wang, Hongfang Liu and Ming Huang	91
<i>LAMAR at ArchEHR-QA 2025: Clinically Aligned LLM-Generated Few-Shot Learning for EHR-Grounded Patient Question Answering</i>	
Seksan Yoadsanit, Nopporn Lekuthai, Watcharitpol Sermsrisuwan and Titipat Achakulvisut	96
<i>Neural at ArchEHR-QA 2025: Agentic Prompt Optimization for Evidence-Grounded Clinical Question Answering</i>	
Sai Prasanna Teja Reddy Bogireddy, Abrar Majeedi, Viswanath Gajjala, Zhuoyan Xu, Siddhant Rai and Vaishnav Potlapalli	104
<i>UIC at ArchEHR-QA 2025: Tri-Step Pipeline for Reliable Grounded Medical Question Answering</i>	
Mohammad Arvan, Anuj Gautam, Mohan Zalake and Karl M. Kochendorfer	110

<i>DMIS Lab at ArchEHR-QA 2025: Evidence-Grounded Answer Generation for EHR-based QA via a Multi-Agent Framework</i>	
Hyeon Hwang, Hyeongsoon Hwang, Jongmyung Jung, Jaehoon Yun, Minju Song, Yein Park, Dain Kim, Taewhoo Lee, Jiwoong Sohn, Chanwoong Yoon, Sihyeon Park, Jiwoo Lee, Heechul Yang and Jaewoo Kang	118
<i>CogStack-KCL-UCL at ArchEHR-QA 2025: Investigating Hybrid LLM Approaches for Grounded Clinical Question Answering</i>	
Shubham Agarwal, Thomas Searle, Kawsar Noor and Richard Dobson	126
<i>SzegedAI at ArchEHR-QA 2025: Combining LLMs with traditional methods for grounded question answering</i>	
Soma Nagy, Bálint Nyerges, Zsombor Kispéter, Gábor Tóth, András Szlúka, Gábor Kőrösi, Zsolt Szántó and Richárd Farkas	136
<i>LIMICS at ArchEHR-QA 2025: Prompting LLMs Beats Fine-Tuned Embeddings</i>	
Adam Remaki, Armand Violle, Vikram Natraj, Étienne Guével and Akram Redjda	150
<i>razreshili at ArchEHR-QA 2025: Contrastive Fine-Tuning for Retrieval-Augmented Biomedical QA</i>	
Arina Zemchyk	160
<i>DKITNLP at ArchEHR-QA 2025: A Retrieval Augmented LLM Pipeline for Evidence-Based Patient Question Answering</i>	
Provia Kadusabe, Abhishek Kaushik and Fiona Lawless	165
<i>AEHRC at BioLaySumm 2025: Leveraging T5 for Lay Summarisation of Radiology Reports</i>	
Wenjun Zhang, Shekhar Chandra, Bevan Koopman, Jason Dowling and Aaron Nicolson	171
<i>MetninOzU at BioLaySumm2025: Text Summarization with Reverse Data Augmentation and Injecting Salient Sentences</i>	
Egecan Evgin, Ilknur Karadeniz and Olcay Taner Yıldız	179
<i>Shared Task at Biolaysumm2025 : Extract then summarize approach Augmented with UMLS based Definition Retrieval for Lay Summary generation.</i>	
Aaradhya Gupta and Parameswari Krishnamurthy	185
<i>RainCityNLP at BioLaySumm2025: Extract then Summarize at Home</i>	
Jen Wilson, Michael Pollack, Rachel Edwards, Avery Bellamy and Helen Salgi	190
<i>TLPIQ at BioLaySumm: Hide and Seq, a FLAN-T5 Model for Biomedical Summarization</i>	
Melody Bechler, Carly Crowther, Emily Luedke, Natasha Schimka and Ibrahim Sharaf	196
<i>LaySummX at BioLaySumm: Retrieval-Augmented Fine-Tuning for Biomedical Lay Summarization Using Abstracts and Retrieved Full-Text Context</i>	
Fan Lin and Dezhi Yu	202
<i>5cNLP at BioLaySumm2025: Prompts, Retrieval, and Multimodal Fusion</i>	
Juan Antonio Lossio-Ventura, Callum Chan, Arshitha Basavaraj, Hugo Alatrística-Salas, Francisco Pereira and Diana Inkpen	215
<i>MIRAGES at BioLaySumm2025: The Impact of Search Terms and Data Curation for Biomedical Lay Summarization</i>	
Benjamin Pong, J u - H u i Chen, Jonathan Jiang, Abimael Jimenez and Melody Vahadi	232
<i>SUWMIT at BioLaySumm2025: Instruction-based Summarization with Contrastive Decoding</i>	
Priyam Basu, Jose Cols, Daniel Jarvis, Yongsin Park and Daniel Rodabaugh	240

<i>BDA-UC3M @ BioLaySumm: Efficient Lay Summarization with Small-Scale SoTA LLMs</i> Ilyass Ramzi and Isabel Bedmar	249
<i>KHU_LDI at BioLaySumm2025: Fine-tuning and Refinement for Lay Radiology Report Generation</i> Nur Alya Dania Binti Moriazi and Mujeen Sung	256
<i>CUTN_Bio at BioLaySumm: Multi-Task Prompt Tuning with External Knowledge and Readability adaptation for Layman Summarization</i> Bhuvaneswari Sivagnanam, Rivo Krishnu C H, Princi Chauhan and Saranya Rajiakodi	269
<i>Team XSZ at BioLaySumm2025: Section-Wise Summarization, Retrieval-Augmented LLM, and Reinforcement Learning Fine-Tuning for Lay Summaries</i> Pengcheng Xu, Sicheng Shen, Jieli Zhou and Hongyi Xin	275
<i>VeReaFine: Iterative Verification Reasoning Refinement RAG for Hallucination-Resistant on Open-Ended Clinical QA</i> Pakawat Phasook, Rapepong Pitjaroonpong, Jiramet Kinchagawat, Amrest Chinkamol, Tossaporn Saengja, Kiartnarin Udomlapsakul, Jitkapat Sawatphol and Piyalitt Ittichaiwong	281